Functional genomics of Crassulacean acid metabolism in the monocot biomass feedstock crop *Agave sisalana*

PHAITUN BUPPHADA

UNIVERSITY OF LIVERPOOL

Thesis submitted in accordance with the requirements of the

University of Liverpool for the degree of Doctor in Philosophy

October 2015

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor, Dr. James Hartwell, for his great advice and support throughout my PhD. I would also like to thank my second supervisor, Dr. Meriel Jones, who always helped me to solve many problems. Advice given by them has been a great help in fulfilling my PhD project.

I would like to offer my special thanks to the following people in the Lab G who always gave me a hand with their lab work experiences when I needed, postdoctoral researchers: Dr. Susanna Boxall and Dr. Louisa Dever, laboratory technicians: Nirja Kadu and Jean Woods, PhD students and lab colleagues: Jade Waller and Jack Davies as well as other people in the lab. I would also like to thank CGR bioinformaticians: Dr. Yongxiang Fang, Dr. Xuan Liu and Dr. Luca Lenzi who assisted me with bioinformatics works.

I would like to extend my thanks to my scholarship provider, the Agricultural Research Development Agency (ARDA), Thailand, who gave me an unforgettable opportunity to come to explore the scientific world at the University of Liverpool and to fulfil my PhD study.

Last but not least I wish to thank my beloved family members, relatives and friends in Thailand and Liverpool who always gave me moral support throughout my difficult times and shared our fun but valuable life experiences together.

Abstract

Certain Crassulacean acid metabolism (CAM) crops have been recognised as having great potential for the production of renewable biomass for bioenergy production from seasonally dry lands. The work described in this thesis sought to investigate the functional genomics of CAM development and light/ dark regulation in the obligate CAM species *Agave sisalana*. Semi-quantitative RT-PCR analysis was employed to study the regulation of CAM genes in leaf tissues. The transcript levels of the CAM genes phosphoenolpyruvate carboxylase (*AsPPC*) and pyruvate orthophosphate dikinase (*AsPPDK*) were highest in the mature tip, lower in the young, expanding base, and very low to undetectable in the most basal white tissue of the youngest fully expanded leaf from ~3-month-old plants. The PEPC kinase gene (*AsPPCK*) did not show a clear pattern of differential regulation of its transcript level between the leaf tip and base.

CO_2 exchange measurements, immuno-blotting of known CAM proteins and malate measurements further confirmed CAM induction in the leaf tip. Furthermore, this is the first report of a circadian rhythm of CO_2 fixation in a monocot CAM species. The phosphorylated form of PEPC was only detected in the leaf tip in the dark. Sucrose was highest in the leaf tip, and showed strong light/ dark regulation and clear evidence for circadian clock control. A putative sucrose metabolism-related gene, cell wall invertase (*As_cwINV*), exhibited strong light/ dark regulation and a robust circadian rhythm in the leaf tip.

De novo transcriptome assembly using Illumina RNA-sequencing data totalling ~90 Gbp was generated from light and dark samples of the white basal, pale green basal, and dark green tip sections of the youngest fully expanded leaf sampled in the light (2 h before dusk) and dark (2 h before dawn). Differential expression analysis identified novel CAM-induced transcription factor genes *AsNAC* (c566713_g1), *AsWRKY* (c571790_g2), and *AsPLATZ* (c541787_g1), which exhibited a robust pattern of both light/ dark regulation and circadian clock control, which was established using Q-RT-PCR analysis.

Overall, this study provides a high quality whole transcriptome assembly and quantitative analysis resource underpinning future functional genomics studies of CAM in *A. sisalana.* The CAM-induced and circadian clock controlled transcription factors identified in this study could also be investigated further through generating stable transgenic RNA interference lines or other approaches to determine their functions. This study has also proposed a novel CAM pathway showing leaf development and light/ dark regulation of CAM genes, including the fructan metabolism pathway, thereby providing a better understanding of how fructan might be synthesised and accumulated and turned over to supply part of the PEP required for nocturnal CO_2 fixation, in addition to the utilisation of sucrose by CAM in the *A. sisalana* leaf.

Table of Contents

List of Figures

List of Tables

Glossary of Abbreviations

µl	microliters
AP2	APETALA 2
ATP	adenosine triphosphate
cwINV	CELL WALL INVERTASE
CAM	Crassulacean acid metabolism
CCA1	CIRCADIAN CLOCK ASSOCIATED 1
CDS	coding DNA sequence
DE	differentially expressed
dH2O	distilled water
EDTA	ethylene diamine tetraacetic acid
EtOH	ethanol
DMSO	Dimethyl sulfoxide
FDR	false discovery rate
FEXH	FRUCTAN EXOHYDROLASE
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
G6PDH	glucose-6-phosphate dehydrogenase
GOT	glutamate-oxaloacetate transaminase
KNOX1	KNOTTED1-LIKE HOMEOBOX1
h	hour
LD	Light/ Dark
LL	Light/Light (continuous light)
M	molar
MDH	malate dehydrogenase
ml	millilitres
NAD	nicotinamide adenine dinucleotide
ORF	open reading frame
PEP	PHOSPHOE*NOL*PYRUVATE
PEPC	PHOSPHOE*NOL*PYRUVATE CARBOXYLASE
PGI	phosphoglucoisomerase
PMSF	phenylmethanesulfonylfluoride or phenylmethylsulfonyl fluoride
PPCK	PHOSPHO*ENOL*PYRUVATE CARBOXYLASE KINASE
PPDK	PYRUVATE ORTHOPHOSPHATE DIKINASE
qRT-PCR	quantitative reverse transcription polymerase chain reaction

RNA-seq	RNA-sequencing
RO H_2O	reverse osmosis water
rpm	revolutions per minute
s	second
TPM	transcripts per million
UBQ10	POLYUBIQUITIN 10
WUE	water use efficiency

Chapter 1

Introduction

1.1 Crassulacean acid metabolism (CAM)

1.1.1 Background, Evolution and Taxonomic Distribution of CAM

Crassulacean acid metabolism (CAM) is an adaptation of photosynthetic carbon dioxide (CO_2) fixation that is believed to have evolved from the ancestral C_3 form of photosynthetic CO_2 fixation in a diverse range of higher plant species. The term 'Crassulacean type of acid metabolism' was published for the first time in the New Phytologist in (Bennet-Clark, 1933) and was coined to describe the daily fluctuations in leaf acid levels that were discovered in the first CAM plants. This first report of CAM was made using a species in the genus *Sedum,* which is a member of the family Crassulaceae, hence the derivation of the 'Crassulacean' in CAM. The nocturnal acid accumulation of succulent plants was also studied in various species within the Cactaceae during this period (Thomas, 1949).

As for the C_4 adaptation of photosynthesis, CAM is a carbon concentrating mechanism (CCM) modified from the ancestral C_3 photosynthetic CO_2 fixation system with an effective biochemical CO_2-pump which increases the CO_2 concentration around the main photosynthetic carbon-fixing enzyme, ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) in the Calvin-Benson cycle. By pumping or concentrating CO_2 around the RuBisCO inside the chloroplasts of each photosynthetic cell, both of these CCMs enhance the capacity for the carboxylase reaction over the oxygenase, and thus, both C_4 and CAM minimise photorespiration, the potentially wasteful side-reaction of photosynthesis that results from

RuBisCO oxygenase activity (Singh *et al.*, 2014). C_4 photosynthesis enhances CO_2 fixation by means of cell-specific spatial separation of primary and secondary CO_2 fixation into the mesophyll and bundle-sheath respectively. By contrast, the CAM CCM occurs in each mesophyll or photosynthetic chlorenchyma cell of the leaf, but there is a temporal separation between CO_2 fixation to vacuolar malate in the dark period and malate decarboxylation and subsequent CO_2 refixation via RuBisCO in the light period (Cockburn, 1985). Strong CAM species fix CO_2 primarily during the dark period, which generally means that their stomata, through which they lose water, are only open when the atmospheric temperature is relatively low, and the humidity is relatively high, and thus, evapotranspiration is minimised relative to C_3 and C_4 plants that open their stomata in the light period. CAM plants close their stomata during the hot, dry light period as they perform secondary CO_2 fixation via RuBisCO using CO_2 released from malate accumulated the night before. This light period stomatal closure prevents evapotranspiration, and so, water loss is considerably lower than a C_3 or C_4 plant. These changes in leaf physiology and metabolism improve the water use efficiency (WUE) of photosynthetic carbon assimilation by 6- to 10-fold relative to C_3 species (Nobel, 1996; Borland *et al.*, 2009). This WUE advantage of CAM correlates well with the fact that the majority of CAM species have evolved in and naturally inhabit desert and semi-arid/ seasonally dry environments (Black and Osmond, 2003).

Although CAM is believed to be the result of adaptation and evolution in drought tolerant plants, it is present in a diverse spectrum of species and life-forms. This brings great taxonomic complexity to this trait, which in turn highlights the need for critical thought when generalising about the CAM pathway (Dodd *et al.*, 2002). Tracking the multiple independent origins of CAM during plant evolution is a challenging scientific endeavour. There has often been a heavy focus on WUE in desert and arid environments being a key selective factor driving evolutionary selection towards CAM (Holtum *et al.,* 1999). However, this photosynthetic adaptation has

also recently been found in aquatic plants, including the fern ally *Isoetes,* whose evolutionary origins can be dated back into the Triassic period, around 100 million years earlier than the occurrence of CAM in terrestrial plants (Holtum *et al.*, 1999; Keeley, 2014). In the terrestrial angiosperms (flowering plants), CAM has evolved many times and has been estimated to be present in approximately 6-7 % of all higher plant species, outnumbering the estimated number of C_4 species by 2-fold (Winter and Smith, 1996; Crayn *et al.*, 2004; Holtum *et al.*, 2007).

CAM plants are believed to have diverged from C_3 ancestors at some point in the Miocene period, possibly due to a decline in the atmospheric CO_2 concentration during that period (Raven and Spicer, 1996). The available evidence to date supports the suggestion that the evolutionary route to CAM occurred via C_3–CAM intermediates before the arrival of full CAM (Pilon-Smits *et al.*, 1996). This occurred alongside the evolution of other specialised adaptations to dry environments in many CAM species (e.g. leaves reduced to spines in cacti), suggesting that the photosynthetic plasticity of CAM often coincided with dramatic speciation events. In several cases, this resulted in rapid speciation and radiation of CAM species within their increasingly semi-arid and arid region (Lüttge, 1996; Kluge *et al.*, 2001).

CAM has been detected in 5 taxonomic classes including at least 35 different families and more than 400 genera of both monocotyledonous and dicotyledonous plants (Winter and Smith, 1996; Winter *et al.*, 2015). The current best estimate for the number of CAM species is 16,000, but this is still likely to be underestimate due to the complexity of the experimental data required to characterise and confirm the presence of CAM, combined with the growing number of surveys that include large numbers of CAM representatives (Dodd *et al.*, 2002).

Many CAM species are classified as facultative (or inducible) species. In these species, CAM can be induced in response to temperature, drought, salinity and/ or high irradiance stress,

and also ageing and photoperiod (Winter, 1985; Winter and Gademann, 1991; Adams *et al.*, 1998; Broetto *et al.*, 2002; Lüttge, 2002; Hurst *et al.*, 2004). Facultative CAM species tend to perform ancestral C_3 photosynthesis under favourable conditions (for example well-watered, and intermediate temperature and light). When facultative species experience a stressful change in their environments (for example the soil drying associated with drought), they exhibit an immediate response by switching to CAM. A facultative CAM pathway has been observed in annual plant species that inhabit seasonally arid regions (Winter and Holtum, 2011), and also in a number of tropical and sub-tropical tree species within the genus *Clusia* (Holtum *et al.*, 1999).

In addition to facultative CAM, a large number of CAM species are categorised as displaying constitutive (or obligate) CAM photosynthesis. These species perform CAM throughout their whole life cycle regardless of the prevailing environmental conditions. Constitutive CAM is found in various desert succulent plants including the cacti and agaves (Smith *et al.*, 1997), but is also found in members of the Crassulaceae, including in the genus *Kalanchoë*, and in a range of tropical and sub-tropical orchids and bromeliads; many of which are epiphytes (Holtum *et al.*, 1999).

The plant species used as the main study species in this work, *Agave sisalana,* is a monocotyledonous, succulent CAM species that belongs to the Agavoideae, a subfamily of monocot flowering plants in the family Asparagaceae, order Asparagales. This subfamily has previously been categorised as a separate family in its own right, namely the Agavaceae, by some authors (Chase *et al.*, 2009). *Agave* is the largest genus within the sub-family Agavoideae. Previous estimates placed the number of species at around 166 but this has recently been revised upwards to a current estimate of approximately 208 species (Good-Avila *et al.*, 2006). The number of *Agave* species are believed to undergone rapid expansion and

radiation around 6 to 8 million years ago (mya; during the late Miocene and early Pliocene period), coinciding with a rise in the frequency and severity of drought in central Mexico (Good-Avila *et al.*, 2006; García Mendoza, 2007). A further key speciation event is estimated to have occurred around 2.5 to 3 mya (late Pliocene period), coincident with the spread of nectarivorous bats, which today are the major pollinators of many extant *Agave* species (Matiz *et al.*, 2013).

Agaves are native to southern and western regions of North America, and central and tropical parts of South America. Most of them are succulents with thick and fleshy leaves forming a rosette, which for some species can reach huge proportions up to several metres across. Each leaf end is usually protected by a sharp spike at its tip, plus leaves generally possess spiny edge along the entire length of the leaves (Gentry, 2004). One of the most economically important parts of the Agave plant is the central heart-bud (piña), a plump stem base that is rich in non-structural carbohydrate, and which has been used by native people as a source of sugar and alcohol for hundreds if not thousands of years. A key modern use of the agave piña from the species *Agave tequilana* cv. Weber Azul (the so-called "Blue Agave") is the production of the popular distilled alcoholic drink known as tequila.

A. sisalana is known by the common name 'sisal' due to the fact that it is grown as a source of the renewable and sustainable cellulose fibre product 'sisal'. Sisal fibre is used for the production of agricultural baler twine and marine ropes, as well as rugs, carpets and dart boards, but more recently has found new uses as a sustainable strengthening agent in modern composites used for the manufacture of products such as car door panels. Sisal has been grown commercially for fibre in a number of key countries and regions of the world including Indonesia, the Philippines, Brazil, and East Africa (Gentry, 2004). Brazil is currently by far the largest producer in the world, although even in Brazil the sisal industry still remains largely a

cottage industry. Aside from its existing commercial exploitation for sisal fibre production, *A. sisalana* is also believed to hold great potential as a renewable and sustainable source of biomass for application such as biofuels and renewable platform chemicals for industrial manufacturing (Borland *et al.*, 2009; Corbin *et al.*, 2015). Researchers in these areas are now becoming increasingly interested in the valuable and favourable characteristics of *A. sisalana* and other existing crop species of Agave due to this potential for new uses for a crop which can grow productively and sustainably in semi-arid and seasonally drought prone agricultural land where most major food crop species would struggle to achieve a respectable yield. Nevertheless, relatively few studies have been reported for the crop species of Agave and this is particularly true for *A. sisalana* (see more details in Section 1.2).

1.1.2 Physiology and Biochemistry of CAM

The daily 24 h cycle of the CAM metabolic adaptation of photosynthesis has been divided into four key phases by Osmond, (1978), which help with understanding the different aspects of cellular biochemistry and physiology that are occurring at different times. These four CAM phases are presented in Figure 1.2. The daily cycle of CAM is outlined in the metabolic pathway diagram shown in Figure 1.1, and includes several physiological and biochemical activities. Phase I starts at night when CAM plants open their stomata. Atmospheric CO_2 entering the leaf air spaces is hydrated by carbonic anhydrase (CA) resulting in the formation of bicarbonate (HCO_3^-). The activity of CA is the first step in carbon sequestration in CAM plants, and its function also helps to prevent the loss of CO_2 back into the atmosphere, facilitating the supply of CO_2 (Tiwari *et al.*, 2005). Phosphoenolpyruvate carboxylase (PEPC) is activated during the dark period through phosphorylation by phosphoenolpyruvate carboxylase kinase (PPCK) (Hartwell *et al.*, 1999). CAM uses PEPC to capture free atmospheric CO_2 by combining phosphoenolpyruvate (PEP) and HCO_3^-, a product from the action of CA on

CO_2, to generate 4-carbon organic compound, oxaloacetate (OAA) which is then rapidly converted to malate by malate dehydrogenase (NAD(P)-MDH) (Ting, 1985). Malate is transported into the vacuole through a voltage-gated, inward-rectifying anion channel and accumulated as malic acid (Hafke *et al.*, 2003). The malate channel activity at the vacuolar tonoplast membrane may be encoded by an aluminium-activated malate transporter (ALMT) gene, although direct evidence for this in a CAM species is yet to be reported (Kovermann *et al.*, 2007; Borland *et al.*, 2009). The nocturnal accumulation of large quantities of malic acid in the vacuole requires energisation via the high concentration of H^+ (protons) imported into the vacuole by the vacuolar H^+-ATPase and/or H^+-PP_iase (Bartholomew *et al.*, 1996; Tsiantis *et al.*, 1996). This pumping of protons into the vacuole during the dark period generates a proton motive force (PMF) that drives the active transport of malate into the vacuole and its internal dissociation to malic acid in the presence of the abundant hydrogen ions (Ward *et al.*, 2009; Martinoia *et al.*, 2007). The uptake and fixation of CO_2 and accumulation of malic acid occur continuously for most of the dark period, leading to the vacuole being swollen and full of malic acid such that the concentration can reach levels up to ~200 mM by dawn (Borland *et al.*, 2009). The vacuole of CAM mesophyll or chlorenchyma cells can reach up to 95 % of the cell volume when it is in this 'malate-filled' state at dawn (Steudle *et al.*, 1980). This malate accumulating activity results in the noticeably high level of acidity in CAM leaves at dawn with the cell sap pH level descending as low as pH 2.9 (Franco *et al.*, 1990). In the few hours before dawn, PEPC kinase is degraded, resulting in PEPC becoming dephosphorylated by the constitutively active protein phosphatase type 2A that dephosphorylates PEPC. This dephosphorylation renders PEPC up to 10-times more sensitive to feedback inhibition by its allosteric inhibitor malate (Nimmo *et al.*, 1987).

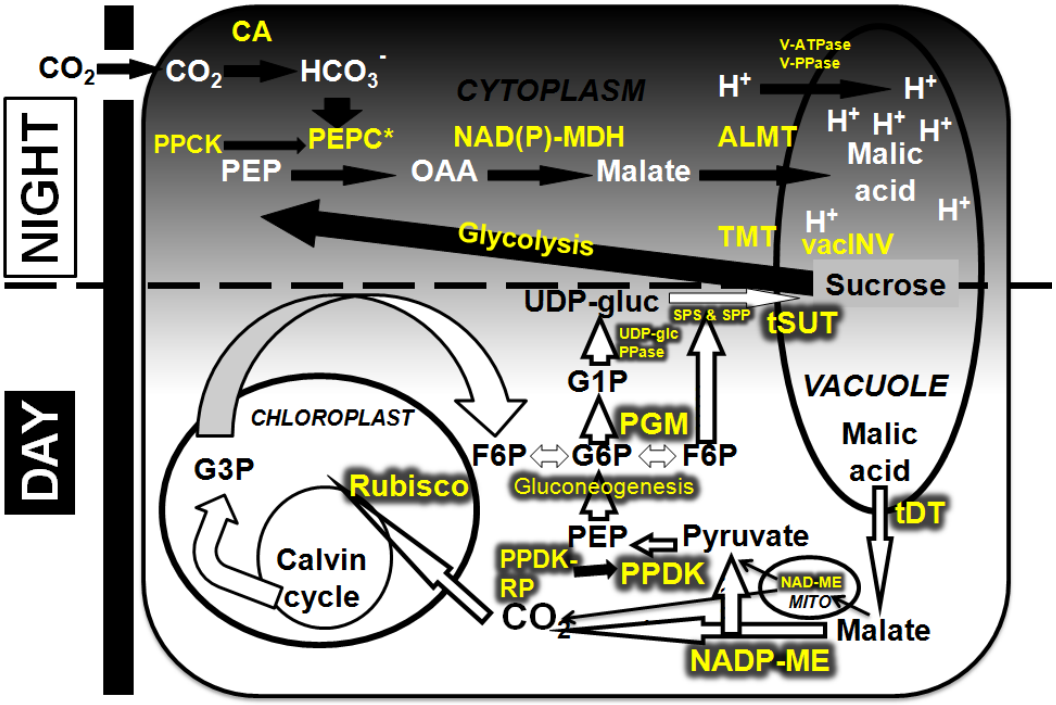


Figure 1.1 Proposed CAM pathway in *A. sisalana*.

A proposed diagram of the CAM pathway in *A. sisalana,* modified from Borland *et al.*, (2009). Filled arrows represent dark reactions whilst empty arrows represent light reactions. The dashed line running across the centre divides the reactions into dark at the top and light at the bottom. The thick black line to the left of the diagram represents the leaf epidermis, with the gap at the top of the line representing a stomatal pore. The enzymes and proteins that are most likely to catalyse the reactions and transport the metabolites across the various sub-cellular compartmental membranes are highlighted in yellow.

During the day-time, the gas exchange profile of CAM plants progresses through phases II, III, and IV respectively (Figure 1.2), with each phase in the gas exchange corresponding to distinct stages in the progression through light period malate decarboxylation and secondary CO_2 re-fixation (Figure 1.1). The malic acid accumulated from the dark period is transported out of the vacuole into the cytosol, possibly by a tonoplast dicarboxylate transporter (tDT) (Emmerlich *et al.*, 2003; Hurth *et al.*, 2005). Due to the convergent and/ or parallel evolution of CAM in many independent taxonomic lineages, malate decarboxylation in the cytosol can proceed via either mitochondrial NAD- and/or cytosolic/ plastidic NADP-malic enzyme (ME), or a combination of cytosolic MDH and cytosolic PEP carboxykinase (PEPCK) (Black *et al.*, 1996; Christopher and Holtum, 1996). The exact pathway of malate decarboxylation varies between individual CAM

species and lineages, consistent with the multiple independent origins for the evolution of CAM that are known (Winter and Smith, 1996). Plants in the family Crassulaceae have commonly been found to employ cytosolic NADP-ME for malate decarboxylation, although recent transgenic evidence gained through the use of gene specific RNAi silencing of mitochondrial NAD-ME in *Kalanchoë fedtschenkoi* revealed that NAD-ME was the major malate decarboxylation enzyme in the light in this member of the Crassulaceae (Cockburn, 1985). Many members of the monocot lineage Bromeliaceae have been reported to use PEPCK for malate decarboxylation in the light (Cockburn, 1985; Dever *et al.*, 2015). In Agaves, like many CAM species and lineages, the exact pathway of malate decarboxylation in the light (via NADP-ME, NAD-ME, or PEPCK) has yet to be elucidated in detail, although several species have been suggested to use both NADP-ME and NAD-ME based on enzyme activity measurements (Cushman and Bohnert, 1997; Christopher and Holtum, 1996). Malate decarboxylation yields CO_2 and also pyruvate or PEP which is subsequently recycled via gluconeogenesis to regenerate a pool of leaf storage carbohydrates that are available during the following dark period for breakdown to PEP for PEPC. Whether pyruvate or PEP is synthesised depends on which of three decarboxylating enzymes are involved. In NAD-ME and NADP-ME based CAM species such as *K. fedtschenkoi* and *M. crystallinum*, malate decarboxylation yields pyruvate and CO_2. This CO_2 is re-fixed via RuBisCO in the Calvin-Benson cycle and pyruvate is converted to PEP by pyruvate, orthophosphate dikinase (PPDK) (Evans and Wood, 1968). This PEP together with glyceraldehyde 3-phosphate (G3P), a product from Calvin-Benson cycle, is metabolised through gluconeogenesis to produce starch, soluble sugars or fructans, depending on the species. In Agaves, sucrose is believed to be the storage carbohydrate which is remobilised at night to provide the PEP for nocturnally primary CO_2 fixation by PEPC (Christopher and Holtum, 1996). In PEPCK-type CAM species such as *Ananas comosus, Aloe vera, Hoya carnosa* and *Neoregelia carolinae*, malic acid is first converted to OAA by MDH,

which is active in the reverse direction relative to its role in the dark during CO_2 fixation, and the OAA is decarboxylated by PEPCK yielding PEP and CO_2 and PPDK is not required (Antony *et al.*, 2008).

Malate decarboxylation during the light period in CAM species creates a high concentration of CO_2 around RuBisCO, which favours the carboxylase activity over the oxygenase (Holtum *et al.*, 1999). This results in a dramatic increase in efficiency of CO_2 fixation and decline of photorespiration. The high internal concentration of CO_2 is also hypothesised to mediate stomatal closure (Cockburn *et al.*, 1979), which then leads to reduced water loss and the maintenance of the high internal partial pressure of CO_2 inside the leaf throughout much of the light period; until malate is exhausted. It has been estimated that the internal inorganic carbon abundance inside a CAM leaf or stem during the light period can reach 60-fold greater than the ambient CO_2 concentration due to malate decarboxylation behind closed stomata (Lüttge, 2002).

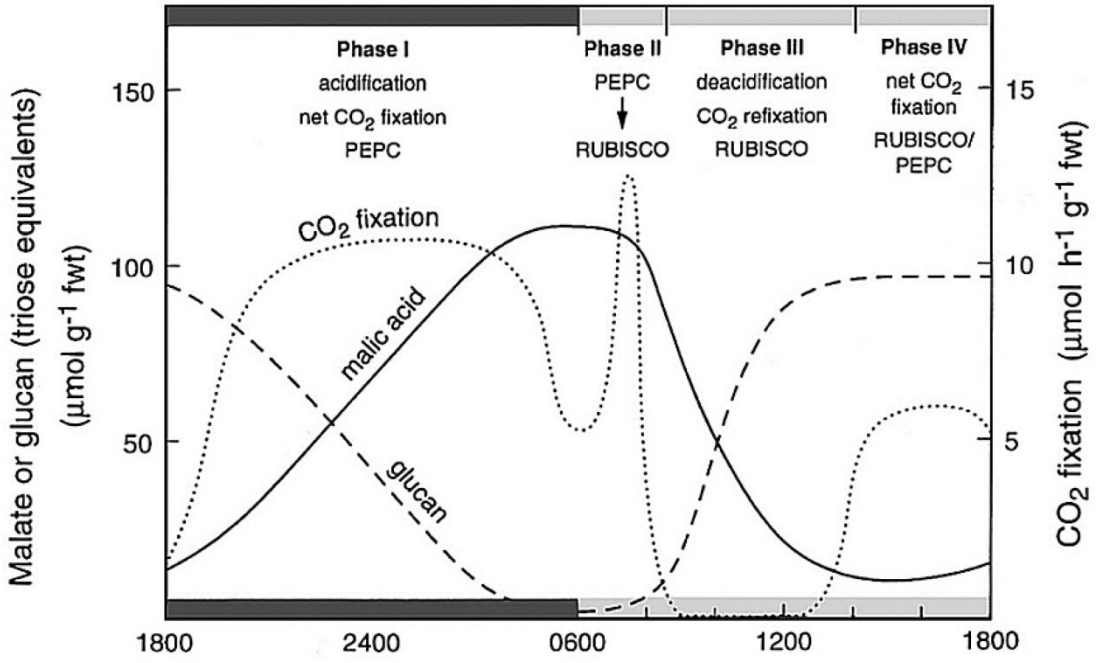


Figure 1.2 The daily cycle of CAM showing the four phases of the daily CO_2 fixation cycle as defined by Osmond, (1978).

Generalised schematic representation of the four phases of CAM CO_2 fixation, alongside the accumulation and turnover of malic acid, and the reciprocal changes in the levels of glucans or other carbohydrates. Each phase is defined in relation to the timing of the main biochemical processes occurring during the 24 CAM cycle. For example, both carboxylation and malic acid decarboxylation, as well as the activities of both PEPC and RuBisCO. The dark bar indicates the dark period (1800 - 0600 h) and grey bar indicates the light period (0600 - 1800 h) along the x-axis. The diagram is reproduced from Black and Osmond, (2003).

Overall, the CAM metabolic phases can be described as follows: phase I and III are the two main primary and secondary CO_2 fixing phases of CAM. I and III are interconnected by the transitional phases II and IV. Phase I defines the period of CO_2 fixation and acidification in the dark, and is initiated by the activation of PEPC and the opening of stomata. These factors facilitate the fixation of atmospheric CO_2 into malic acid which is then transported into the vacuole where it accumulates throughout the dark period. Phase III coincides with the main period of malate decarboxylation. Stomata close and Rubisco is active in the refixation of the CO_2 released from malate. The transitional phases II and IV provide additional net atmospheric CO_2 uptake at dawn and in the late afternoon, respectively, although pronounced phase II and

IV activity is only detected in well-watered plants, and for younger leaves or cladodes of constitutive CAM species. Older leaves and cladodes and partially drought-stressed CAM tissues tend to rely almost exclusively on phases I and III. When phase II does occur under favourable conditions, both PEPC and RuBisCO are active for the first few hours of the light period, and stomata remain open, allowing a short and rapid period of atmospheric CO_2 fixation to occur via both the C_3 and C_4 enzymes (Griffiths *et al.*, 1990). During phase IV, which occurs in young tissues of a well-watered CAM plant in the late afternoon, stomata are able to re-open once all of the malate has been decarboxylated and thus the internal CO_2 concentration inside the leaf has declined. This allows the direct fixation of atmospheric CO_2 via RuBisCo, which is exactly the same as the photosynthetic CO_2 fixation occurring throughout the light period in a C_3 plant (Griffiths *et al.*, 1990). Phase IV completes the daily 24 h cycle of CAM and the whole cycle starts again after dusk when nocturnal phase I recommences. The duration and magnitude of each CAM phase displays huge plasticity and variation; both with environmental conditions and between CAM species. For example, in the constitutive CAM dicot *K. fedtschenkoi*, the duration and magnitude of nocturnal phase I and light period phase III have been shown to expand as a leaf matures and develops (Borland *et al.*, 2009).

1.1.3 Temporal regulation and circadian clock control of CAM

The C_4 adaptation of photosynthetic CO_2 fixation relies on spatial compartmentation of distinct biochemical steps into distinct specialised cell-types (mesophyll and bundle-sheath) in order to achieve enhanced CO_2 fixation in the light via PEPC and RuBisCo. By contrast, CAM occurs in each mesophyll or chlorenchyma cell and relies on strict temporal regulation of the biochemical steps in order to achieve nocturnal, primary CO_2 fixation and malic acid accumulation, and subsequent light period malate decarboxylation and secondary CO_2 refixation, without futile cycling occurring between the two processes (Cockburn, 1985). Strict

temporal control of CAM is essential for the efficient operation of the pathway (Hartwell, 2006). Efficient timing of the key steps of the CAM pathway is achieved through tight coupling of CAM enzymes and transporters and their regulators to the endogenous circadian clock.

Daily on-going zeitgebers (German term for 'time givers'), such as light/ dark cycles and temperature cycles (changes of temperature levels), are the inputs that entrain the central circadian oscillator of plants to coordinate output pathways to their surrounding environments (Salomé and McClung, 2005). Circadian clock systems are one of the subset of biological rhythms and fulfil three key rules (McClung, 2006). Firstly, they are endogenous and self-sustaining rhythms which persist even in free-running conditions without the entrainment driven by environmental input such as light and temperature cycles (Salomé and McClung, 2005). Secondly, they have a period of approximately 24 h (approximate time to complete one cycle) (Dunlap *et al.*, 2004). Thirdly, they are temperature compensated. Among rhythms in living organisms, they have a period of oscillation that shows a high degree of temperature compensation, such that the period does not change significantly across the normal range of temperatures experienced by that organism (Wilkins, 1992). The mechanism of temperature compensation is supposed to prevent the circadian oscillator from the direct influence of changes in cellular metabolism which accelerate with the increase of temperature (McClung, 2006). The study of CO_2 fixation rhythms in CAM *K. fedtschenkoi* and *K. daigremontiana* supports these three parameters for a bona fide circadian rhythm, revealing that CO_2 fixation persists with rhythms in free-running conditions for up to 10 days over a range of ambient temperatures (Wilkins, 1992).

Plants are mostly sessile photosynthetic organisms, and their circadian clock manages and controls many aspects of their biology in coordination with the rhythmic surrounding environment generated through the daily rotation of the earth on its axis and the resulting 24

h light/ dark and hot/ cold cycle. Possession of a central circadian oscillator allows plants to optimise their biology in relation to the light/ dark cycle (McClung, 2006; Hartwell, 2005). The plant circadian clock has been proved to control a wide range of fundamental biochemical and physiological processes including cell and hypocotyl elongation (Jouve *et al.*, 1998; Dowson-Day and Millar, 1999), flowering time (Yanovsky and Kay, 2003; Imaizumi and Kay, 2006), leaf movement (Engelmann and Johnsson, 1998), photosynthesis, gene expression, metabolic pathways, gas exchange (Hennessey and Field, 1991), stomatal opening and closing (Webb, 1998) and CO_2 assimilation in CAM plants (Nimmo, 2000). In *A. thaliana*, an important study found that the plant clock optimised the efficiency of photosynthesis through appropriate matching of the circadian clock period with the light/ dark cycle of the external environment (Dodd *et al.*, 2005). This led to improved growth performance in plants grown under light/ dark cycles that matched to the period of their endogenous circadian clock, whereas clock mutants whose endogenous clock period differed from 24 h performed less well than the wild type unless they were grown under artificial light/ dark cycles that matched the period of their endogenous clock (Dodd *et al.*, 2005). It has also been found that as many as one third of genes in *A. thaliana* are under circadian clock control (Covington *et al.*, 2008). In *M. crystallinum*, several CAM-associated genes only become clock-controlled when the plant is induced to perform CAM by salt-stress treatment (Boxall *et al.*, 2005; Cushman *et al.*, 2008). These findings suggests that circadian control is important in for the optimal functioning of the CAM pathway (Hartwell, 2005).

In plant biology, CAM was one of the first physiological traits to be found to be under the control of the circadian clock. Measurement of net CO_2 exchange in CAM leaves of *Bryophyllum* (*Kalanchoë*) *fedtschenkoi* under constant dark and temperature conditions revealed that the dark period phased fixation of CO_2 in this CAM plant was able to persist under constant conditions (Wilkins, 1959). Further understanding of the molecular and

biochemical basis for the circadian control of CAM has been achieved through the study of the circadian regulation of key enzymes within the CAM pathway, especially PEPC. At night, PEPC is phosphorylated on an invariant, N-terminal serine residue by PEPC kinase (PPCK), making PEPC up to 10-times less sensitive to feedback inhibition by malate, which is the ultimate product of its activity during the dark period (Nimmo *et al.*, 1984; Nimmo *et al.*, 1986; Carter *et al.*, 1991). A study by Carter *et al.*, (1991) in *K. fedtschenkoi* leaves revealed that the activity of PPCK was subject to regulation by an endogenous circadian clock. The activity of PPCK in CAM leaves of *K. fedtschenkoi* increased in the dark and declined during the light period. A highly correlated light/ dark pattern of regulation was also observed for the phosphorylation state of PEPC and the apparent K_i of PEPC for malate, which represent the consequences of PEPC phosphorylation by PPCK. Furthermore, a subsequent study using *in vitro* translation of total RNA isolated from CAM leaves of *K. fedtschenkoi* followed by an *in vitro* assay for the activity of PPCK in the translation products revealed that *PPCK* translatable RNA increased in the dark under the control of the circadian clock (Hartwell *et al.*, 1996). In addition, it was found that the nocturnal increase in the level of *PPCK* translatable RNA could be prevented by transcription and translation inhibitors (Hartwell *et al.*, 1996). The novel method for the *in vitro* translation of CAM leaf RNA followed by assay of PPCK activity was subsequently adapted to allow the cloning of the gene encoding *PPCK*. A cDNA library generated using RNA from CAM leaves of *K. fedtschenkoi* sampled in the middle of the dark period was transcribed into mRNA *in vitro* and the transcripts were subjected to *in vitro* translation and PPCK activity assays. A process of several rounds of screening pools and sub-pools of the cDNA library ultimately led to the identification of a single clone that encoded a full length PPCK, which generated extremely high levels of PPCK activity when transcribed and translated *in vitro* (Hartwell *et al.*, 1999). This report of the cloning of the first *PPCK* gene from *K. fedtschenkoi* was followed shortly afterwards by the cloning of the CAM PPCK gene from the inducible CAM

species *M. crystallinum* (Taybi *et al.*, 2000). The *PPCK* gene was the first CAM-associated clock-controlled gene (CCG). The transcript level of *PPCK* was under circadian clock control in mature leaves of *K. fedtschenkoi*, peaking in the middle of dark period (Hartwell *et al.*, 1999). In the closely related species, *K. daigremontiana,* experiments that prevented nocturnal CO_2 fixation and malate accumulation by encapsulating whole CAM leaves in pure nitrogen gas during the dark period revealed that the circadian control of PPCK degradation at dawn could be overriden, possibly due to the dramatic decrease in the accumulation of malate during the dark period (Borland *et al.*, 1999). When leaves were returned to normal air at dawn, there was a short burst of CO_2 fixation and malate accumulation which correlated with a decrease in *PPCK* translatable RNA and activity, and the phosphorylation state of PEPC also declined (Borland *et al.*, 1999).

The circadian system has been categorised into three essential components: inputs of signals from the environment, central oscillator or 'clock' which sets the time, and rhythmic outputs generated by the clock (Somers *et al.*, 1998). Elements of the circadian clock can be defined based on their expression pattern timing in relation to morning-, daytime- and evening-phased genes (Greenham and McClung, 2015). In *A. thaliana*, the latest model of circadian clock, proposed by Greenham and McClung, (2015), consists of a sequence of several positive and negative inter-connected auto-regulatory feedback loops of genes and their products which transcriptionally and post-transcriptionally regulate one another in a reciprocating and rhythmic way, with also an involvement of post-translational modification and protein turnover (Harmer, 2009) (Figure 1.3). At dawn, the morning-phased genes *CIRCADIAN CLOCK-ASSOCIATED 1* (*CCA1*) and *LATE-ELONGATED HYPOCOTYL* (*LHY*) positively regulate expression of two of the *TIMING OF CAB EXPRESSION 1* (*TOC1)* paralogs, daytime-expressed genes *PSEUDO-RESPONSE REGULATOR 9* (*PRR9*) and *PRR7*, which then interact with *PRR5* and act as negative-feedback regulators of *CCA1* and *LHY*. The PRRs also inhibits *REVEILLE 8* (*RVE8*),

which activates the expression of evening-phased genes: *TOC1*, *EARLY FLOWERING 4* (*ELF4*) and *LUX ARRHYTHMO* (*LUX*). *LUX, ELF3* and *ELF4* together create the evening complex (EC), a multi-protein complex that negatively regulates the expression of daytime-expressed genes *PRR7* and *PRR9*. *TOC1* also negatively regulates the expression of morning-phased genes *CCA1*, *LHY*, *PRR9* and *PRR7*. Apart from transcriptional regulation, there is also post-translational regulation involved in the network. PRR5 and TOC1 protein abundances are regulated via the complex-dependent degradation including SKP, CULLIN and F box, mediated by the F box protein ZEITLUPE (ZTL), which is stabilized by GIGANTEA (GI) (Figure 1.3) (Greenham and McClung, 2015).

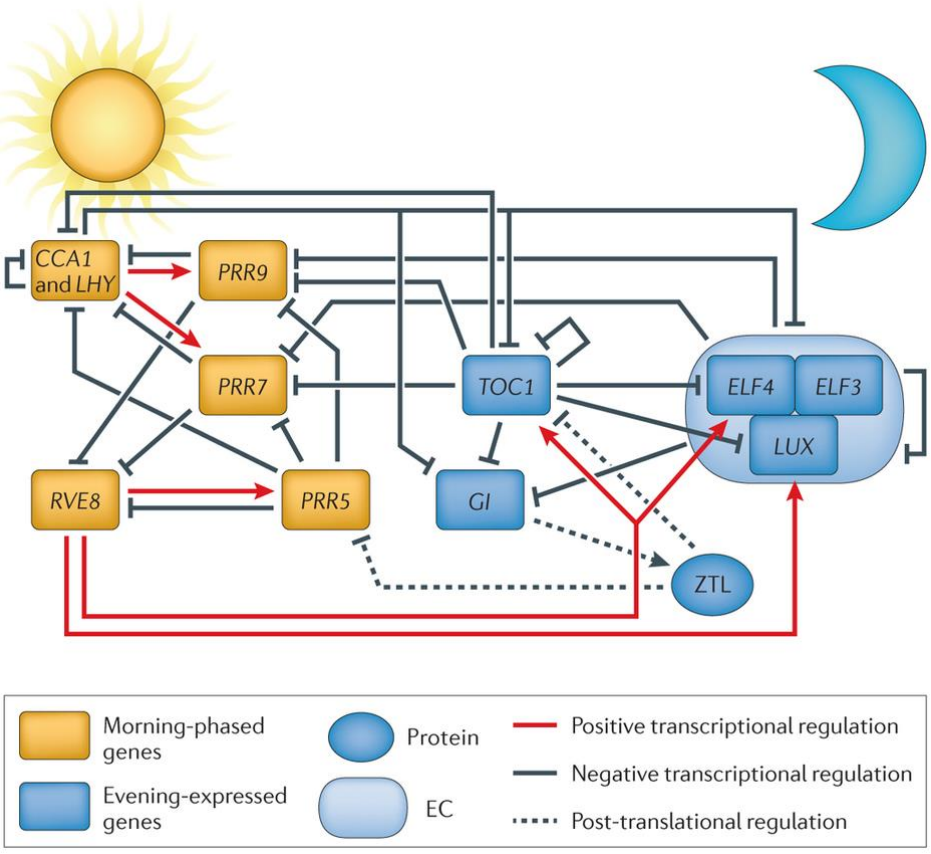


Figure 1.3 The proposed model of the multi-loop molecular circadian clock in *A. thaliana*, by Greenham and McClung, (2015).

The genes involved in the morning loop (yellow) include *CCA1* and *LHY, PPR5, PPR7, PPR9 and RVE8.* The genes involved in the evening loop include *TOC1, GI, ELF3, ELF4* and *LUX. LUX, ELF3* and *ELF4* together form the evening complex (EC). The image was reprinted from: Greenham and McClung, (2015).

1.2 The importance of CAM plants for humans: *Agave sisalana*, as a potential biomass source for renewable biofuels and platform chemicals for industry

As already mentioned, CAM plants close their stomata to perform secondary CO_2 fixation during the hot, dry light period and this minimises evapotranspirational water loss during the most stressful part of the day. This results in a significant increase in the water use efficiency (WUE) of carbon assimilation in CAM plants, such that they are able to achieve WUE values up to 3-fold greater than C_4 species, and between 6- to 10-fold greater than C_3 species (Nobel, 1996; Borland *et al.*, 2009; Borland *et al.*, 2015). This is a key trait that helps CAM species to be considered as some of the most drought tolerant plants. Possession of the C_4 photosynthetic system has been shown to help certain C_4 crops such as maize or sorghum to maintain high photosynthetic rates at higher temperatures that would take C_3 photosynthesis beyond its CO_2 compensation point. However, even though C_4 species generally achieve elevated WUE relative to C_3 species, sustaining their potential high productivities relies on water availability, which in practical terms often leads to a requirement for supplementary irrigation (Stewart, 2015). The ability to conserve water during CO_2 assimilation in the dark period has often been considered to be the most significant functional characteristic of CAM plants (Winter *et al.*, 2005). A range of CAM crop species are employed for agriculture in arid, semi-arid and seasonally drought prone regions. CAM crops have been found on average to require approximately 20% of the water required by C_3 or C_4 crops, and yet can achieve similar above-

ground biomass yields (Borland *et al.*, 2009). Others have estimated that their biomass production per unit water used is 6-fold higher than C_3 and 2-fold higher than C_4 crops (Winter *et al.*, 2005). The most commonly cultivated CAM species, such as *Ananas comosus* (pineapple), *Opuntia ficus-indica* (prickly-pear), *A. sisalana* (sisal), and *A. tequilana* (used for production of the Mexican alcoholic beverage 'tequila') have been reported to achieve yields at nearly maximum capacity in regions where annual rainfall is not sufficient for the cultivation of a number of C_3 and C_4 crops (Borland *et al.*, 2009). High-productivity Agave species are already grown on a commercial scale for the production of alcoholic beverages (e.g. tequila and mescal produced from *A. tequilana* and *A. salmiana* in Mexico), natural fibres (e.g. sisal fibre from *A. sisalana* or *A. fourcroydes*) and soluble fibre for functional foods in the form of fructans (inulin). It has been estimated that more than 100 M and 130 M ha of land are commercially utilised for cultivation for fibre and beverage production respectively from Agave species (Nobel, 2010). Corbin *et al.*, (2015) determined the major biomass and bioenergy associated constituents of *A. americana* and *A. tequilana* leaves, and reported that the dry Agave leaf fiber contained 47–50% (w/w) of crystalline cellulose content and 16–22% (w/w) of non-cellulosic polysaccharides content, while the lignin content was low (9–13% w/w). Soluble sugars, cellulose, non-cellulosic polysaccharides, lignin, acetate, protein and minerals together accounted for 85–95% of the total dry mass of whole Agave leaves. Extracted juice from the Agave leaves also accounted for 69% of the fresh weight and was found to be rich in glucose and fructose (Corbin *et al.*, (2015). The total sugar content measured from the Agave piña (heart of the plant) ranges from 12 % to 28 % in fresh weight. Agave piñas and leaves are also observed to have high percentage of water content (Yan *et al.*, 2011), which could help reduce volume of water required for biofuel processing.

Agave leaves are succulent and fibre-rich allowing for the persistence of CO_2 fixation and other essential biochemical activities through the prolonged dry periods (Lüttge, 2004; Matiz *et al.*,

2013). Agave roots can adapt in response to drying soils by shrinking themselves to minimise their water loss (Nobel and Sanderson, 1984). They can also form fine roots rapidly following a short-lived rain event to allow quick uptake of water (North and Nobel, 1998). Other attributes of Agaves as suitable crops include their long-lived perennial habit (Nobel, 1994; Bowers *et al.*, 1995; Gentry, 2004), semelparous flowering (Eguiarte *et al.*, 2000), natural tendency to grow on stony, infertile lands (Shreve, 1942), and easiness of vegetative propagation (Arizaga and Ezcurra, 1995).

Agave sisalana, the central study species of the work described in this thesis, has great potential as a source of biomass for biofuel production. Favourable characteristics of CAM plants to be considered for the cultivation of bioenergy crops in semi-arid lands can be listed as shown in Table 1.1. *A. sisalana* has a number of favourable and potential end uses relative to other biofuel feedstock crops, including high water-use efficiency, tolerance to drought, high biomass yields with high concentrations of non-structural carbohydrate (less energy required for conversion to fuels) and high cellulosic content (43 %; Davis *et al.*, (2011)), low lignin content (an obstacle for lignocellulosic fuels synthesis; 15 %; Davis *et al.*, (2011)), and a number of high value co-products such as fructans, which have a wide variety of uses in the food, medical and bio-ethanol industries (Maldonado-Sanchez, 2009; Borland *et al.*, 2009; Stewart, 2015).

Table 1.1 Favourable growth attributes of CAM plants for cultivation as a bioenergy crop in semi-arid areas, table taken from Borland *et al.*, (2009).

Trait	Example	Comment
High water-use efficiency	5–16 mmol CO_2 per mol H_2O on an annual basis	Typically 4- to 10-fold higher than C_3 plants
High drought tolerance	Can grow in areas with as little as 25 mm $year^{-1}$ precipitation	Tissues can tolerate up to 90% loss of water content (cacti)
Tolerance of high temperatures	Up to 70 °C, based on 50% loss of cell viability after 1 h; can survive exposure to 74 °C	Typically upper limit of 50–55 °C in C_3 plants
Tolerance of high PPFD	Can tolerate >1000 μmol m^{-2} s^{-1} (or >40 mol m^{-2} d^{-1}) without photoinhibition	Generally more tolerant of high PPFD than agronomically important C_3 plants
Tolerance of UV-B radiation	Only 1% incident UV-B transmitted through epidermis of *Yucca filamentosa* (Agavaceae)	Generally thick epidermis and high foliar concentrations of phenolics in CAM plants
Entire shoot surface typically photosynthetic	Whole shoot photosynthetic in both leaf- and stem-succulent species; limited bark formation even on stems of arborescent cacti	Many C_3 species deciduous (shedding photosynthetic organs for part of year) or woody (limited stem photosynthesis)
High shoot:root ratio and harvest index	Shoot:root ratio as high as 10:1; above-ground biomass readily harvested	
High resistance to herbivores	Effective physical defences (stem succulents) and chemical defences (leaf succulents)	
High content of non-structural carbohydrate	Especially monocotyledons (∼20% dry weight); ready conversion of soluble sugars to bioethanol	
Low lignin content	Weak secondary thickening and lack of true wood formation	

Due to the current rapid climate change, shrinking oil reserves and predicted food shortages, there is an increasingly urgent and pressing need to develop new biomass-based carbon sources for the production of liquid transportation fuels which are carbon neutral but do not impact on food security (Aosaar *et al.*, 2012). The existing first generation biofuel feedstocks struggle with a number of limitations. For example, most of biomass feedstocks exploited for biofuels production nowadays have to compete with the use for food industry for human consumption (Ragauskas *et al.*, 2006; Somerville *et al.*, 2010). The use of corn (maize, *Zea mays*) for the production of bioethanol eliminates corn from the food supply chain. This conflict between the use of food crops for food and/ or biofuels has been widely cited as at

least part of the reason for the 2008 surge in world food prices and the associated food-riots in various developing countries (Eugenio *et al.*, 2009). It is generally accepted in the scientific literature that the world cannot afford to use food crops as biofuel feedstocks due to the fact that many of the world's major food crops, such as rice, wheat, and maize, have reached a yield plateau, and food shortages are expected to become an increasing problem in the next 10-20 years as the population increases from the 7 billion today to the expected > 9 billion by 2050 (Lee *et al.*, 2011).

A challenging future has been predicted for global agriculture, particularly in terms of it managing to continue to achieve increased yields to keep track with climate change and population growth. A range of factors are leading to these concerns, including a diminishing supply of often degrading fertile arable land suited to high-productivity growth of productive food crops, the global human population is predicted to continue to rise, currently at 7.2 billion, but with estimates suggesting it will rise by 33 - 71% by 2100 (Gerland *et al.*, 2014), plus climate change models suggest that there will be an increase in the frequency and intensity of severe drought events, bringing an associated decline in soil moisture (Dai, 2013; Cook *et al.*, 2014). This 'perfect-storm' of combined pressures on global agricultural production and food security are likely to strengthen the need for the efficient cultivation of semi-arid and seasonally dry regions with productive crops such as the existing CAM crops already mentioned. CAM crops also provide an opportunity to bring degraded or abandoned agricultural land back into productive use without putting pressure on existing areas of highly productive and fertile agricultural land that are the mainstay of current food production. Furthermore, there is currently a growing realisation that biofuel feedstocks must be produced using land that is not suited to food production. This is likely to be feasible as the area of marginal lands has been reported to be increasing all over the world due to the over-exploitation of existing productive agricultural land (Kendall and Pimentel, 1994). It has been

estimated that there is a large area of abandoned agricultural land around the world that could be reclaimed and exploited through the cultivation of CAM crops for bioenergy or the production of platform chemicals for industry (Davis *et al.*, 2011).

In the light of these major challenges, there is a need to develop new non-food crops as biofuel feedstocks. The various usages of Agave, in particular *A. sisalana* as a subject of this study, together with its distinctive adaptations to environmental changes and all the potentials described above, make *A. sisalana* a proper model CAM crop that exhibits promising potential as an alternative resource of biomass for bioenergy (Stewart, 2015). Whilst it will not be a 'cure-all panacea' for bioenergy and the food security crisis, *A. sisalana* and other CAM crops have the potential to make a valuable and sustainable contribution to humanity's response to the challenges facing global food and energy production.

1.3 Functional genomics of CAM in *Agave sisalana*

The facultative, dicot CAM model plant, *M. crystallinum,* which shifts from C_3 to CAM in saline or xeric conditions, is a well-established model species for the study of the molecular, biochemical and physiological differences between the C_3 and CAM photosynthetic pathways (Adams *et al.*, 1998). The model obligate, dicot CAM species, *K. fedtschenkoi* exhibit a developmental transition from C_3 to CAM even when grown under well-watered conditions (Jones, 1975; Hartwell *et al.*, 1999; Dever *et al.*, 2015). Younger leaves (1st and 2nd leaf pair down the stem from the shoot apex) perform CO_2 fixation mostly during the light period, implying that they undergo C_3 photosynthesis (Borland *et al.*, 2009). By contrast, fully mature and developed leaves (4th and 5th pair and older) fix CO_2 mostly in the dark period, thus performing full CAM (Borland *et al.*, 2009). Moreover, the C_3-to-CAM transition can also be characterized by substantial changes in both the expression level of key CAM genes, and the associated activity of the encoded enzymes and transporters and their regulators enzymes.

Known CAM genes of *K. fedtschenkoi* are under developmental control and are also fully expressed in the 4th to 5th leaf pair and in other older leaves (Hartwell *et al.*, 1999; Borland *et al.*, 2009).

Whilst *M. crystallinum* and *K. fedtschenkoi* are manageable and well-studied model systems that will greatly enhance understanding of CAM, these species are distantly related to CAM species like Agave that exhibits great potentials as biomass source with relatively high yields for bioethanol production (Borland *et al.*, 2009). In comparison to the relative wealth of published scientific literature about CAM photosynthesis in other CAM species, Agave has been much less intensively studied and reported on regarding its CAM photosynthetic biochemistry, regulation and plasticity (Matiz *et al.*, 2013), and has not been the subject of widespread breeding despite the fact that several Agave species have a number of attractive agronomic characteristics as mentioned above. A relatively small amount of molecular-level work has been carried out to improve Agaves, apart from for *A. tequilana* (Stewart, 2015). Thus far, detailed transcriptome sequencing studies have been published for only two Agave species, *A. tequilana* and *A. deserti* (Gross *et al.*, 2013), although other Agave species including *A. americana* and *A. victoriae-reginae* are also the subject of detailed transcriptome sequencing work to enable gene discovery in these species (Avila de Dios *et al.*, 2015; Yang *et al.*, 2015). Avila de Dios *et al.*, (2015) reported that the transcriptome sequencing data of *A. victoriae-reginae* and *A. striata* was generated but has not been published yet. Furthermore, the complete genome sequence of a diploid *A. tequilana* accession is underway (Yang *et al.*, 2015). Considering the relative dearth of knowledge and research on the molecular and biochemical aspects of CAM in Agave, further research is needed to investigate whether CAM photosynthesis in Agave plants is actually influenced by the plant and/or leaf developmental stage. Seemingly, the regulatory processes controlling CAM expression in Agaves remain obscure (Matiz *et al.*, 2013).

Lujan *et al.*, (2009) performed an interesting study in which they demonstrated that the spike (the youngest folded leaves at the centre of the plant) of *A. tequilana* had the highest level of thermotolerance. This section of the plant also exhibited the highest density of stomata and high levels of HSPs (heat-shock proteins). Taking into consideration the fact that this is the youngest part of the plant, these young tissues are most likely to show lower levels of CAM expression; just as young leaves in an obligate CAM species such as *K. fedtschenkoi* display C_3 photosynthesis. The uncertainty about whether young Agave leaves exhibit lower expression of CAM is still a question to be answered, however some observations appear to suggest a promising impact of leaf morphological stages on the level of CAM expression in Agaves (Matiz *et al.*, 2013). In *A. deserti*, there has been a report showing that late-afternoon CO_2 fixation (Phase IV) declines as the young plants gain maturation, nearly disappearing in full-grown plants of this species (Nobel, 1985), which possibly indicates that the transitional development to full CAM is dependent on nocturnal CO_2 fixation (Phase I). Holtum *et al.*, (1999) also stated that young photosynthetic leaf tissues of constitutive CAM plants such as the Agaves are frequently C_3, while CAM is always found at mature leaves. However, the magnitude of the CAM phases is generally responsive to abiotic stresses such as water availability, high-light and high temperatures.

A. sisalana has been found to undergo a developmental transition from C_3 to CAM during leaf development. Based on the findings described earlier, comparing C_3 with CAM photosynthesis in *A. sisalana* could be achieved relatively simply by comparison between samples of young and old tissues of the same leaves from the same plant. In the Hartwell lab, an initial study of the developmental control of CAM CO_2 fixation and the developmental induction of CAM-related genes such as *PPC, PPCK* and *PPDK* was performed (Boxall, Waller and Hartwell, unpublished). The results revealed that CAM is fully induced and developed in the leaf tip (mature green leaf tissue) of *A. sisalana*. In addition, CAM CO_2 fixation in the dark declined

down the leaf such that the leaf base (pale green, younger tissues) used mainly C_3 photosynthesis fixing CO_2 only during the light period. The longitudinal transition observed in the gas exchange patterns was also reflected in increases in the transcript abundance of key genes encoding the major CAM enzymes.

Borland *et al.*, (2009) performed a meta-analysis of the physiological, biochemical and morphological features associated with CAM used by Agave, with a particular focus on its potential for bioenergy production. Comprehensive transcriptome sequencing (RNA-seq), proteomic, metabolomic and further physiological analysis of Agave would be a significant further step to more deeply understand CAM in this important CAM crop species. Furthermore, a detailed study of the regulation of photosynthetic genes in relation to physiological and environmental conditions would provide valuable insights into the developmental and environmental control of CAM in Agave (Hartsock and Nobel, 1976; Nobel and Valenzuela, 1987; Silvera *et al.*, 2010). Generating reference Agave leaf transcriptomes would allow investigations at molecular level in order to further understand the correlated gene expression underlying CAM in Agave (Gross *et al.*, 2013).

The quantification and identification of mRNA in various plant tissues has been a key area of recent rapid breakthroughs in plant biology; largely driven by the arrival of increasingly productive high-throughput DNA sequencing systems that generate massive amounts of sequence information quickly and for a relatively low price. Through the past decade, there have been two major theoretically different methods used to study gene expression profiling (Tyagi, 2000). The first approach relies on microarrays whereby cDNA is hybridized to an array containing complementary oligonucleotide probes that correspond to genes of interest, and the mRNA abundance is determined based on intensity of hybridization associating with probes (Schena *et al.*, 1995). The second method includes a popular SAGE approach

(Velculescu *et al.*, 1995) and recently a massively parallel signature sequencing (MPSS) (Brenner *et al.*, 2000). These methods are based on the sequencing of cDNA fragments and calculating a number of times a target fragment has been detected. However, these approaches still have comparatively restricted sequencing method that does not vigorously identify rare mRNAs (Wang, 2007). On the other hand, next-generation sequencing approaches provide considerably higher sequencing throughput at a much lower cost per sample (Morozova and Marra, 2008). Such technology permits progressively deeper transcriptome sequencing making the identification of more transcripts realistic expectation today (Haas and Zody, 2010).

Comprehensive analysis of the transcriptome is crucial for understanding the functional components in the genome and identifying the molecular elements of cells and tissues, and their development and response(s) to the environment. RNA-seq has a number of core advantages over other existing methods although it is still a relatively new and developing field (Wang *et al.*, 2009). The advantages include the fact that RNA-seq short reads provide information about the connectivity of two or more exons. RNA-seq has no limitation of detecting transcripts that correspond with the existing genomic content. Moreover, RNA-seq is able to discover sequence variations (e.g. single nucleotide polymorphisms or SNPs) within the transcribed regions of a genome (Cloonan *et al.*, 2008; Morin *et al.*, 2008). These advantageous features make RNA-seq a suitable tool for exploring complex transcriptomes. Another benefit of RNA-seq in comparison to microarrays is that there is no need to perform amplification and cloning, thus fewer RNA samples are required. High reproducibility for both biological and technical replication has also been observed in the outputs of RNA-Seq (Cloonan *et al.*, 2008; Nagalakshmi *et al.*, 2008). Importantly, RNA-Seq has been proved to be very precise for gene expression quantification. Lastly, there is no limitation for quantification

which associates with the amount of data sequenced, thus RNA-seq subsequently enables expression levels to be determined for a huge dynamic range of transcripts.

Overall, with all of these advantages, RNA-Seq has become one of the key methods allowing biologists to discover novel genes in non-model species where genome sequencing has not been completed or not currently available. Based on sequencing techniques, RNA-Seq is the first method that permits the whole transcriptome to be explored in a high-throughput and quantitative way. This technology enables the study of digital gene expression (DGE) levels at a genome-wide scale and it has already been used extensively by hundreds if not thousands of researchers (Haas and Zody, 2010; Wang *et al.*, 2009). Furthermore, rapid further progress in the development of DGE techniques is predicted, such that RNA-seq is likely to out-compete microarray approaches for several functional genomics studies in the coming years. A key data analysis step during any RNA-seq or microarray experimental work-flow relates to the calculation of the probability as to whether or not read counts for a gene transcript of interest are significantly different amongst the studied experimental conditions (Robinson *et al.*, 2010). A typical feature of RNA-seq is that it commonly sequences and creates enormous volumes of data (Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008). Thus, a differential expression analysis tool with powerful statistical capabilities and an appropriate error model is required to deduce differential regulation of individual genes within such huge datasets; in order to determine and account for the data variability across the dynamic range (Anders and Huber, 2010).

Agave plants are known to have a large genome, with an estimated size of approximately 4 Gbp or larger depending on the species (Palomino *et al.*, 2003). Agave is believed to be of paleopolyploid-origin, and thus the genome commonly contains a large amount of gene duplication (McKain *et al.*, 2012). This taken together with the known large proportion of repetitive elements in Agave genomes (Bousios *et al.*, 2007), make complete genome assembly

for Agaves a challenging task. Gross *et al.*, (2013) performed the first transcriptomics study of Agave, and their assemblies represented the first published Agave transcriptome. They generated *de novo* transcriptomes of two Agave species, *A. tequilana* and *A. deserti*, reporting at least 35,000 protein-coding genes for each species (Gross *et al.*, 2013). They also reported a comparative, quantitative transcriptome profiling experiment carried out with *A. deserti*, providing an overview of the molecular and physiological functions that change between different segments of the leaf. This study revealed that the developmental progression from the young basal portion of a monocot leaf to the mature tip or apical portion of the leaf is conserved widely across monocot evolution. In the C_4 species, maize (*Zea mays*), RNA-seq transcriptome analysis identified the differentially expressed genes that changed in transcript abundance along the developmental gradient of a young maize leaf (Li *et al.*, 2010). Moreover, Zhou *et al.*, (2012) reported that a pooled-tissue method was efficient for the preparation of sequencing libraries for effective deep sequencing in *A. sisalana,* and thus for the subsequent discovery of novel genes. In the Hartwell lab, an *A. sisalana* RNA-seq experiment using 454 sequencing was carried out previously to begin to determine the CAM transcriptome of *A. sisalana* and its light/dark regulation. This work generated 1.2 million 454 reads with key CAM genes including *PPC, PPCK, MDH, PPDK, V-ATPas*e subunits, *NAD*- and *NADP-ME* having been identified. Genome sequencing of Agave has not been achieved or at least is not available yet. However, the genome of a diploid form of *A. tequilana* is currently being sequenced, assembled and annotated as part of a large International collaborative project in the United States (CAM Biodesign project: http://cambiodesign.org) and is going to be completed in coming months (Hartwell, personal communication; Yang *et al.*, (2015)). When such a reference genome is available, it would greatly facilitate and enhance the study of CAM in *A. sisalana*.

1.4 Fructan metabolism in Agaves

Agave plants have adapted during their evolution to grow in arid and semi-arid regions due to several important characteristics including physiological components of fructan accumulation (Lopez *et al.*, 2003). Fructan metabolism has been suggested to be an evolutionary adaptation of Agave to survive in long periods of drought (Hendry, 1993). Thus, another classic characteristic of Agave plants is fructan metabolism and production. Fructans are composed of long β-fructofuranosyl polymers synthesised from sucrose and stored in vacuoles of succulent parenchymatic cells found in stems and leaves (Mancilla-Margalli and Lopez, 2006). Thus, in Agaves fructans represent a significant vacuolar sink for the products of photosynthesis. The major storage form of non-structural carbohydrates found in Agave are fructans; they comprise 60 % or more of the total water soluble carbohydrate (WSC) (Davis *et al.*, 2011). Fructans are proposed to play an important role to metabolism and development of plants, such as osmoregulation, cryoprotection, and drought tolerance (French, 1989; Ritsema and Smeekens, 2003). In adult Agaves, fructans are a key source of energy used to fuel flowering, which involves the production of a tall flower spike that expands and elongates rapidly over a very short period of time. It has been found that in mature leaves of *A. deserti*, biosynthesis of fructans occurred only in the vascular tissue (Wang and Nobel, 1998). The mechanism whereby fructans are transported to sink tissues, and consequently stored in the vacuole, remains unknown. It is believed that synthesis of fructans occur in the vacuole through the activity of fructosyl transferases that utilise imported sucrose as their substrate (Valluru and Van den Ende, 2008). Agaves produce a variety of fructans (Borland *et al.*, 2009) with a broad variety in structure (Mellado-Mojica and Lopez, 2012). A wide diversity of sugar and fructan content has been found in a variety of Agave species (Vargas-Ponce *et al.*, 2007). Mellado-Mojica and Lopez, (2012) studied fructan Metabolism in *A. tequilana* and found that molecular

structures of fructan develop to be more highly complex when the plants age, especially by the end of their development. During the life cycle of the plants, fructan metabolism showed changes in carbohydrate and fructan contents, fructan degree of polymerization (DP), type, and molecular structure. While some reports have indicated that fructans are not broken down at night to supply PEP for the nocturnal CO_2 assimilation associated with CAM in Agaves (Raveh *et al.*, 1998), other studies suggest that these carbohydrates are possibly used as the major source of nocturnal PEP production (Olivares and Medina, 1990). In addition, Avila de Dios *et al.*, (2015) very recently studied fructan related genes in *A. tequilana, A. deserti, A. victoriae-reginae,* and *A. striata* using transcriptome analysis obtained using the RNA-seq approach. These authors found that they were unable to identify assembled transcripts encoding two key fructan metabolism enzymes in their Agave transcriptome assemblies, namely fructan:fructan 1-fructosyltransferase (1-FFT), an enzyme responsible for chain elongation of inulin-type fructans, and sucrose:fructan 6 fructosyltransferase (6-SFT), a key enzyme for diverting carbon from sucrose to fructan. To date, the carbohydrates that Agaves produce and use, and their metabolism, as well as the impact(s) of environmental factors and leaf and rosette development remains relatively poorly understood (Matiz *et al.*, 2013). Thus, in addition to current interest in developing detailed understanding of the CAM pathway in Agave, the further investigation of fructan metabolism is also an interesting challenge and could enhance the understanding of CAM in Agave.

1.5 PhD project aims

This PhD project sought to understand the molecular basis for high yield and high water-use efficiency in *A. sisalana*. The favourable characteristics and potentials of *A. sisalana* as a biomass source for biofuels, as mentioned above, have raised the interest in more deeply understanding CAM by investigating and identifying CAM genes that are related to Agave

productivity, hence facilitating future improvement of Agave as a potential source of biomass for bioenergy production. The latest high-throughput DNA sequencing technologies were employed to decipher the molecular-genetic basis for the metabolic adaptation of CAM which is one of the key adaptations underlying Agave's high WUE.

The main objective of the project was to investigate the biochemistry and functional genomics of CAM in *A. sisalana* with a particular focus on understanding the molecular signalling pathways that are involved in the coordination of CAM relative to the endogenous circadian clock. A key goal was to achieve a transcriptome-wide view of the genes that *A. sisalana* uses to perform CAM. Overall, the project involved growing a variety of Agaves, selecting species, sampling their transcriptome with high-throughput DNA sequencing, identifying genes that correlate with high productivity and water use efficiency, then characterising the regulation of these genes over the light/ dark cycle and during circadian free running conditions in constant light and temperature. Molecular works included applying quantitative real-time RT-PCR techniques to measure the abundance of target gene transcripts across time course samples. In addition, protein and metabolite analyses were used to investigate the developmental regulation of several CAM-associated phenotypes along the leaf developmental gradient including protein abundances for CAM-associated proteins measured with immunoblotting, the levels of CAM-associated metabolites (malate and sugars), and the CO_2 exchange rhythms of the different leaf segments over the light/ dark cycle. These CAM phenotype measurements defined a robust framework for the extent to which different sections of the *A. sisalana* leaf were performing CAM. These findings thus aided greatly with the interpretation of the RNA-seq data in the light of the associated CAM physiology and biochemistry.

A second important and commercially valuable feature of Agave biomass is the soluble carbohydrates, especially fructans, that Agaves synthesise and store. These fructans make

Agave biomass a good source of carbohydrates for fermentation into ethanol for bioenergy production and for biorefining into other novel renewable platform chemicals such as surfactants. The functional genomics of Agave fructan metabolism and storage/accumulation have not been characterised in detail. Thus, a study of fructan metabolism related genes is the second area of interest for further work. Considering that fructans accumulate more in the leaf base relative to the leaf tip, the existing sampling strategy for the above mentioned CAM focused RNA-seq work also had the potential to identify key genes in the fructan pathways (Wang and Nobel, 1998).

The proposed significance of the project

The major goal of this project was to create genomic, specifically transcriptomic, resources to underpin the genetic improvement of CAM crops in the future through facilitating the generation of knowledge to inform engineering of CAM machinery into C_3/C_4 crops, and/or to help with the development of molecular markers to help with future attempts to improve Agave through selective breeding accelerated through the use of molecular markers.

Chapter 2

Materials and Methods

2.1 Plant

2.1.1 Initial scoping experiment: investigation of the timing and localisation of peak transcript abundance for a range of known CAM-associated genes in A. sisalana

Agave sisalana Perrine plants were originally obtained from a commercial nursery (Agave Nursery/ North and South Succulents, Alfreton, Derbyshire, UK run by Jon Dudek) and subsequently propagated from side shoots in a mix of commercial peat based compost (Sinclair Compost) containing one third perlite plus Osmocote slow-release fertiliser applied at the manufacturer's recommended level. Pot size for young plants was 120 cm^3 and plants were moved on into fresh compost in larger pots (735 cm^3) when they were 6 months old. In years two and three of the project, young *A. sisalana* bulbils were obtained directly from the proprietor of the Agave Nursery (Jon Dudek) and were supplied directly from his nursery in Portugal. These clonal bulbils allowed for the establishment of large developmentally synchronised populations of young *A. sisalana* plants at Liverpool, which allowed for the required biological replication of the experiments described in this thesis.

Plants were grown in a greenhouse under a 16 h light/ 8 h dark cycles maintained using supplementary lighting (minimum ~250 µmoles m^{-2} s^{-1} at plant height; maximum light intensity reached > 2000 µmoles m^{-2} s^{-1} during the summer months) provided by high intensity sodium lamps (Son-T); minimum temperature was maintained at 23˚C with supplementary greenhouse

heating and the lights on and 18˚C in the dark period. Daytime maximum temperatures rarely exceeded 35˚C in the summer months due to computer control of automatic cooling fans and roof vents.

In the first preliminary experiment, one-year-old *A. sisalana* plants cultivated in the greenhouse as described above were destructively sampled at 6 h after the light started into liquid nitrogen as follows: meristem, primordial leaves around the meristem, 4 cm long base and tip of youngest leaves above meristem, 4 cm long base and tip of immature leaves (3rd leaf from core), 4 cm long base and tip of mature leaves (8th leaf from the outer-most leaf of the rosette), 2-3 mm thick of fibrous tissue below meristem and above root, fibrous tissue with root initial attached, rhizome (swollen root), nodes from the stolon left from the mother plant, and inter-nodal sections of this stolon. In addition, two samples were taken from stolons detached from a maturing three-year-old plant. These samples were 4 cm long basal and tip sections from the newly developed stolons. Thus, there were a total of 15 samples.

In other preliminary experiments, approximately 3-year-old, and 1.5-year-old plants propagated from side shoots of two-year-old plants, were placed in a climate-controlled plant growth cabinet (Snijders Microclima MC-1000) under a 12 h light/12 h dark cycle (Light ~450 µmoles m^{-2} s^{-1}, 25˚C, 60 % humidity; dark 15˚C, 70 % humidity). Plant were entrained under there 12:12 LD conditions for at least 2 weeks prior to sampling in order to allow the plants to adjust to the light/ dark cycles. A 2-cm-long section from the leaf base and a 5-cm-long section of the leaf tip were sampled from both plant ages. Leaves were harvested by twisting left and right until their base cracked away from the plant, and then leaf sections were cut and frozen immediately in liquid nitrogen according to different time course experiments as follows:

● *Mature plant light/ dark samples*: the sampling time was at 2 h before dusk (light) and 2 h before dawn (dark) using plants under 12:12 LD cycles. The youngest fully expanded leaf (first

leaf separated fully from the central meristematic cone) and mature leaf from approximately 3-year-old plants were sampled and leaf tip and base collected from both leaf ages, making 8 samples in total.

● *Young plant light/ dark samples:* sample times were same as above in the light and dark and the youngest fully expanded leaf was sampled from 1.5-year-old plants, making 4 samples in total.

● *Young plant, leaves sampled every 4 h*: sampling commenced 2 h after lights-on, and then every 4 h over a 12:12 light dark cycle (6 samplings). The youngest fully expanded leaf was sampled from 1.5-year-old plants and base and tip collected separately, making 12 samples in total.

2.1.2 RNA-seq and metabolic and physiological analysis

11 – week – old *A. sisalana* plants propagated from commercially obtained flower spike bulbils were placed in a controlled light/ dark cycle Snijders growth chamber (MC-100) with a 12 h light/12 h dark cycle, temperature of 25^{o}C in light and 15^{o}C in dark, relative humidity of 60% in light and 70% in dark and light intensity of 450 µmole/m^2/s at plant leaf level. This was done 2 weeks prior to sampling in order to allow the plants to entrain to the 12:12 light/ dark cycles. The plants were mixed around randomly within the growth chamber 3 days prior to sampling in order to mix well and obtain the most equal light intensity across all the plants as the light intensity in the Snijders growth chamber was variable (higher in centre and lower at the edges). After swopping, 21 plants in total were placed and numbered in ordered positions as shown in Figure 2.1.

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21

Figure 2.1 The ordered positions of *A. sisalana* plants placed in the Snijders growth cabinet.

The plants were swopped around 3 days prior to sampling in order to obtain the most equal light intensity across all the plants as a means to reduce edge-effects.

The plants were taken out of the pot and the leaves were peeled off starting from the outer most leaves in order to access the whole leaf. The youngest fully expanded leaf was cut into parts transverse cuts (Figure 2.2). A 2-cm-long section was sampled from the white basal and pale green basal part (numbered 1 and 2 in Figure 2.2), and a 5-cm-long section was sampled from tip of leaf (3 in Figure 2.2). The smallest leaf in the centre of the plant (having peeled away the outer leaves of the central meristematic cone) was also cut in half into an approximately 5-cm-long lower (base; numbered 4 in Figure 2.2) and upper (tip; numbered 5 in Figure 2.2) section (Figure 2.2).

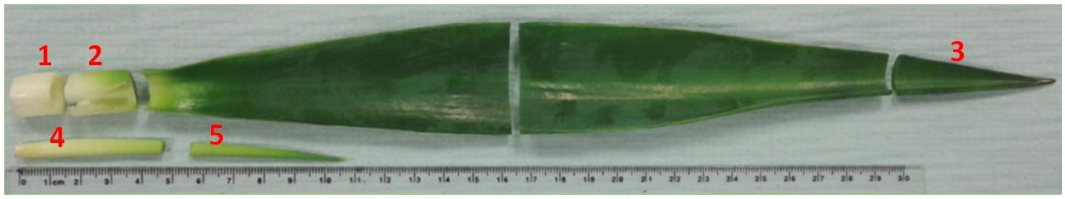


Figure 2.2 Photograph showing the samples collected from different sections of the *A. sisalana* leaves.

The youngest fully expanded leaf was cut into sections for sampling including: 2-cm-long White basal (1) and pale green basal part (Base) (2), and 5-cm-long leaf of Tip (3). The smallest leaf in the centre of the plant (having peeled away the outer leaves of the meristemic cone) was also cut in half into approximately 5-cm-long lower (base) (4) and upper (tip) (5) section.

Leaf sections were cut and frozen immediately in liquid nitrogen according to the specified time courses. The plants were sampled at 2 h after lights-on as the first time point and sampled again at 4 h intervals for 24 h. The last sample was collected at the first time point of the next day (2 h after lights-on). At each sampling time point, 3 plants (3 biological replicates) were sampled. The plants were randomly sampled according to random numbers (Table 2.1) generated using Random Sequence Generator (RANDOM.ORG, http://www.random.org/sequences).

Table 2.1 The random numbers generated using Random Sequence Generator on 2013-01-14 at 10:53:17 UTC.

Sampling Time point	**2 h**	**6 h**	**10 h**	**14 h**	**18 h**	**22 h**	**2 h2**
Plant Number sampled	7	3	12	8	4	9	10
	19	11	6	21	14	15	1
	18	5	20	16	13	2	17

* grey-shaded background indicates the dark sampling time points
* 2 h2 indicates first time point of the next day (2 h after lights-on)

2.1.3 Constant light, temperature and humidity (LL) free-running conditions experiment to test for circadian clock control of genes and metabolite levels

9 – month – old *A.sisalana* plants propagated from commercially obtained flower spike bulbils were fed with liquid fertilizer: dissolving 10 g VITAX VITA FEED 111 in 10 L water. One-month prior to sampling, an equal volume (600 ml) of liquid fertilizer was fed to each plant once in place of the normal watering in order to ensure the plants were not nutrient limited. The plants were then placed in a Snijders MC-1000 growth cabinet under 12 h light/12 h dark cycles, temperature 25˚C in the light and 15˚C in the dark, relative humidity of 60 % in light and 70 % in dark, and light intensity of 450 µmole/m^2/s at plant leaf level. Plants were transferred to the growth cabinet 2 weeks prior to sampling to allow the plants to pre-entrain prior to sampling. The plants were watered regularly and mixed around randomly within the growth chamber 3 days prior to sampling in order to mix well and obtain the most equal light intensity across all the plants as the light intensity varied slightly within the Snijders growth cabinet (higher in centre and lower at the edges). After swopping, 21 plants in total were placed and numbered in ordered positions as shown in Figure 2.1. After 2 weeks of LD pre-entrainment, the light conditions were switched to circadian free-running constant light, temperature and humidity (LL) conditions: continuous light intensity of 100 µmole m^{-2} s^{-1}, 15˚C, 60% humidity. The youngest fully expanded leaf was sampled by cutting the upper half of the whole leaf using a clean sharp razor blade or scalpel. Leaf samples were frozen immediately in liquid nitrogen and stored at -80˚C until use. The length of the whole leaf was approximately 30 cm and the upper half sampled was approximately 15 cm. The leaf of the plants was sampled at 2 h and then every 4 h for 82 h. Each sampling was collected from a single individual plant. Although bioreplicates were not sampled at each time point, circadian biologists often consider that each 24 h cycle under LL conditions is a replicate. As this LL time course covered

3 days, each subsequent cycle was a replicate of the preceding 24 h cycle. The plants were randomly sampled according to random numbers generated using Random Sequence Generator (RANDOM.ORG, http://www.random.org/ sequences) as shown in Table 2.2.

Table 2.2 The random numbers generated using Random Sequence Generator on 2013-10-28 at 11:58:58 UTC.

Sampling time point	2h	6h	10h	14h	18h	22h	26h	30h	34h	38h	42h
Plant number sampled	8	14	20	5	19	3	21	11	17	6	15

Sampling time point	46h	50h	54h	58h	62h	66h	70h	74h	78h	82h
Plant number sampled	7	12	18	16	9	13	4	2	10	1

The overall plant sampling and tissue preparation in various time course experiments can be seen in Table 2.3.

Table 2.3 Overall stages of all experiments carried out in this study giving information on ages of plant, time courses, plant tissues and genes analysed in different experiments.

A. sisalana	Sampling time courses	Sampled parts	Experiments
Initial scoping experiment			
1-year-old	Light	Root - leaf tip	Transcripts level: *AsPPC, AsPPCK, AsPPDK*
3-year-old	Light (2h before dusk) and Dark (2h before dawn)	Tip&Base of young and mature leaves	Transcripts level: *AsPPC, AsPPCK, AsPPDK*
1.5-year-old	Light (2h before dusk) and Dark (2h before dawn)	Tip&Base youngest fully expanded leaf	Transcripts level: *AsPPC, AsPPCK, AsPPDK*
1.5-year-old	Every 4h starting 2h after light for 24h (6 time points in total)	Tip&Base youngest fully expanded leaf	Transcripts level: **CAM:** *AsPPC, AsPPCK, AsPPDK*, *AsNAD-ME, AsNADP-ME, AsNADP-CH, AsALMT* **Clock:** *AsCCA1, AsGI, AsPRR7* **Sugar:** *As_cwINV, AsSPS, AsSUSY, AsFFT, AsFEXH, AsGPPase, AsPGI, AsPGM, AsTST, AsGWD, AsTMT*
Main experiment			
11-week-old (samples prepared for **RNA-seq**)	Every 4h starting 2h after light for 24h. Then repeat 1st time point next day. (7 time points in total)	Tip, Base, Basal white part of youngest fully expanded leaf. And tip & base of smallest leaf in centre.	Transcripts level: *AsPPC, AsPPDK, As_cwINV, AsGI.* Metabolic and physiological analysis
Constant conditions experiment (82 h LL)			
9-month-old	Every 4 h under LL conditions for 82 h	Tip (upper half of the whole leaf)	Transcripts level: *As_cwINV, AsGI AsNAC, AsPPC, AsPLATZ, AsWRKY,* Metabolic and physiological analysis

2.2 Total RNA extraction

Plant samples collected as described in section 2.1 were snap-frozen in liquid nitrogen. Approximately 100 mg of each frozen sample was ground to a fine powder in liquid nitrogen in a pestle and mortar prior to total RNA isolation using the QIAGEN RNeasy® Mini Kit (Qiagen, Germany) according to the manufacturer's instructions with 13.5 µl Polyethylene glycol 20,000 (PEG, 50mg/ml) added to the buffer. This addition of PEG was shown previously to facilitate the recovery of RNA from the leaves of CAM, which have often been found to be challenging for RNA isolation (Boxall and Hartwell, unpublished). The amount of total RNA recovered was determined by measuring the absorbance at 260 nm using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific). The intactness of the total RNA was also determined by visual analysis using denaturing formaldehyde-MOPS agarose gel electrophoresis with 0.5 or 1 µg of each RNA sample used depending on the amount of sample available. These gels were prepared and run according to standard procedures (Bryant and Manning, 1998).

2.3 Semi-quantitative RT-PCR analysis

2.3.1 Primer design

The primers for CAM, circadian clock, sugar-metabolism related and reference/ loading control genes were specifically designed for each gene sequence identified using BLAST searching against an in-house *A.sisalana* 454 transcriptome assembly generated by the Hartwell lab (Boxall, Gregory and Hartwell, unpublished). BLAST searches and primer design were performed using Geneious programme version 5.4 (Drummond *et al.*, 2011).

Table 2.4 A list of primer sequences with annealing temperature for each CAM, circadian clock, sugar-metabolism related and reference genes studied in this work. Primers were designed using Geneious using the built-in Primer3 algorithm.

Gene names	Acronym	Forward Primer	Reverse Primer	Tm (°C)	PCR cycles
CAM genes					
Phosphoenolpyruvate carboxylase	*AsPPC*	TATGGGGACCTCTGACTTGC	TTCTTGGGATCATGCTTTCC	50	27
Phosphoenolpyruvate carboxylase kinase	*AsPPCK*	ACGAGAAGGTGGACATTTGG	TTCCAACAAACTCAAACCACA	51	25
Pyruvate Orthophosphate dikinase	*AsPPDK*	TAGCAATGGGAACCCTGAAC	CTGATACGCAAAGTGGCTGA	51	25
NADP-malic enzyme	*AsNADP-ME*	GGGAGCAGCAGCAGCAGTATTCAG	GAAGCATTCGTGCTTGTTTGCGGG	59	29
NAD-malic enzyme	*AsNAD-ME*	ATGCCTTGCTGCATACATGA	TACGCATCGAATGCCTCTAA	51	25
Chloroplastic NADP-malate dehydrogenease	*AsNADP-MDH*	GGGAGCAGCAGCAGCAGTATTCAG	GAAGCATTCGTGCTTGTTTGCGGG	59	25
Aluminium-activated malate transporter	*AsALMT*	TGAAGTCCATGAGGCAGCCGAAGA	GCTTCAGGTTGGTGGCCGATTGTT	59	26
Circadian clock genes					
Timing of CAB expresstion1	*AsTOC1*	CCAAAAGGGCAGGAACTAGA	ATGAGCACCAAGCACATCAA	51	30
Circadian clock associated 1	*AsCCA1*	ACAGCCAGAACCAAAGAAGCCACCAACG	TGCCAGGAGCAAGCAAAAACGCAGTCCA	62	25
Gigantea	*AsGI*	GATGGGCTCCAATTCTCGTCGCTG	GCAGTCACTTGTGGTGGCATGGTT	58	25
Pseudo-response regulator 7	*AsPRR7*	TGCTGAGTCCCTAAGTCCTGGTGC	AGTGACAGAAGCAACATGCCACGG	58	25
Sugar-metabolism related genes					
Sucrose phosphate synthase	*AsSPS*	CACGGTTGCAGGTTATTCCT	GCTCATGATCTCCCCTGAAC	51	25
Sucrose synthase	*AsSUSY*	TGTTGCCAATGAGTTGGTTG	CACCCGATAGAGTCCAGGAA	51	25
Fructan fructan 1-fructosyltransferase	*AsFFT*	AGGCTGCAATCGAG	GCACTGCAAATGTGC	62	25
Cell wall invertase	*As_cwINV*	CCAACATCGCTCCCCAAAGGCAACCTCC	CTACTGCCTTTGCCGTATTGACGAACGC	62	27
Glucose pyrophosphorylase	*AsGPPase*	GGGGGACTCTTCACACACACAC	AGAAGAAGAGGGACATTGCACCCA	58	25
Phosphoglucose isomerase	*AsPGI*	TGAAAGGGGAGATTGTGAGC	TAGAAGAATCCGCTGGGATG	51	25

Gene names	Acronym	Forward Primer	Reverse Primer	Tm (°C)	PCR cycles
Phosphoglucomutase	*AsPGM*	GCCTAATTGTGGCTCC TTCA	TCAGCACTGGGCAAC TTATG	51	25
Tonoplast sucrose transporter	*AsTST*	TGGAAAAACTCTGTAG AAAGTGG	GCATCACAATTGCCA AGTTCA	51	26
Glucan water dikinase	*AsGWD*	AGCTCCTCATATCCCC GTTC	TGATCAGAAGGATGG CCTTT	52	26
Tonoplast monosaccharide transporter	*AsTMT*	GGGGGTGGATTTTGA TGGCGGTTT	CCCTCTAATATCTGCC GGGGCTGT	59	25
Reference gene					
Polyubiquitin 10	*AsUBQ10*	CATCACCCTGGAGGT GGAGAGCTCGGAC	AGCAATATCCATTCA CAGCCCACCGCGA	63	23

2.3.2 Reverse Transcription PCR and PCR cycles

1 µg total RNA was reverse transcribed to complementary DNA (cDNA) using the QuantiTect® Reverse Transcription Kit (QIAGEN, Germany) according to the manufacturer's instructions with first strand cDNA synthesis incubation time of 45 min. The 20 µL reaction containing the first strand cDNA products was diluted with 80 µL ribonuclease-free water. 1 µL diluted cDNA was then mixed with PCR master mix containing 5 µL REDTaq ready mix (REDTaq® ReadyMix™ PCR Reaction Mix, Sigma-Aldrich, UK), 1 µL forward and reverse primer (final concentration 1 µM) and 2 µL nuclease-free water. Three technical replicates (PCR reactions) were carried out for each sample making the total amount of reactions 3 times greater than the amount of samples. These reactions were amplified simultaneously using a DNA Engine Dyad Peltier thermal cycler (MJ Research) with the following cycling parameters: 95 °C for 2 min, anneal at a primer specific annealing temperature for 30 s, extend at 72 °C for 1 min, denature at 95 °C for 30 seconds. The annealing, extending and denaturing steps were repeated for a number of cycles (Table 2.4). The last cycle ended with incubating at 72 °C for 7 min. The annealing temperature was set as 3 °C below the lowest primer Tm. The number of PCR cycles was optimized for each primer pair to ensure that the PCRs reached an endpoint within the

exponential phase of the amplification of that gene, so that difference in the amount of RNA for that gene in the original RNA samples could be measured.

2.3.3 Gel electrophoresis

PCR products were separated and visualized using a 1 % TAE/agarose gel containing ethidium bromide at a final concentration of 0.1 µg/ml (Ogden and Adams, 1987). 2.5 µl 1000 bp Hyperladder (Bioline, UK) was also loaded into the first lane of each gel to allow estimation of the size of the PCR products. Agarose gels were run in 1×TAE running buffer containing 0.1 µg/ml ethidium bromide at 100V for 30 min using Fisher Scientific FB 300 gel tanks.

2.3.4 Gel image intensity determination for transcript quantification

Gels were imaged using a U: Genius Syngene UV/ visible light gel imaging system (Syngene, Cambridge, UK) with an exposure time of 0.350 set to optimize the image. Images were digitally captured using the Syngene system and analysed using the Metamorph programme (Meta Imaging Series 6.1, Universal Imaging Corp.) to determine the intensity of each gene-specific PCR product band by calculating the integrated intensity from the image. The transcript abundance of CAM, circadian clock and sugar-metabolism related genes, calculated as the average of three technical replicates for each gene, was normalized to the transcript values of a reference gene, Polyubiquitin 10 (UBQ10). This reference gene was tested for its expression stability in an earlier study in 2011 using freely available programmes; namely Best Keeper (Pfaffl *et al.*, 2004) and NormFinder (Andersen *et al.*, 2004) and ranked as one of the best reference genes compared to other candidate genes tested (Bupphada and Hartwell, unpublished results). The normalized quantified band intensities for each PCR product were used to plot graphs to allow relative comparison of the abundance of each transcript across different plant tissues and time courses.

2.4 Q-RT-PCR

2.4.1 Primer design

Primers for qRT-PCR were designed to the sequences of genes of interest, according to Life Technologies™ Real Time PCR handbook (2014 edition). Each gene's sequence was BLAST – searched, using Geneious programme version 5.4 (Drummond *et al.*, 2011), against the Trinity transcriptome assembly generated from the Illumina RNA-seq data produced as part of this project. A pair of forward and reverse primers was designed to the region outside the longest open reading frame (ORF), as the ORF region was more likely to be highly conserved with other homologous and/ or paralogous genes within the *A. sisalana* genome. Where possible, primers were specifically designed to target the 3' untranslated region, as this non-coding region has a tendency to be more highly divergent between closely related members of a gene family. Primers were set to be 18–24 nucleotides in length in order to provide for practical annealing temperatures. The PCR product size was set from 50 to 150 bp, melting temperature from 57˚C to 63 °C with a maximum difference of 3˚C between the Tm's of the two primers. GC content of 50 – 60 % was set in this range in order to ensure that the Tm would be reasonable.

Table 2.5 A list of *A. sisalana* Q-RT-PCR primer sequences used in this study.

Gene names	Acronym : Contig name	Forward Primer	Reverse Primer	Product size (bp)	Tm (°C)
Novel transcription factors (TF)					
No apical meristem TF	*AsNAC* : c566713_g1	GGGAGGATGCGTG TGTCTAT	ATCAGGTTAACCCC CAAGAGA	137	60
WRKY domain TF	*AsWRKY* : c571790_g2	ATGTGGAAATGGCC TACTGG	ACATCGCAGAAGTT GTACGC	59	60
PLATZ TF	*AsPLATZ* : c541787_g1	ATCTCATCTTGGGG CTTCCT	TTAACCATTGGACC CACCAT	146	60
BTB/POZ domain TF	*AsBTB* : c599899_g1	TCAAGCTCAGGATG CAGATG	ACGGACACCACTTT CTTTGC	76	60
Apetala 2 TF	*AsAP2* : c582092_g1	GTAAAGGGGTGCT GCTCAAC	TTCTCCTTCACGAAC CTGCT	124	60
DOF domain, zinc finger TF	*As_zf_DOF* : c534926_g1	TGTGCAAGCTCTGT GGTTTC	CATTGCGGAGAGCT TCAGAT	128	65
Class I Knotted1-like homeobox TF	Class I *AsKNOX1* : c568644_g2	GCGTTGGCCTCAAT TGTACT	GGGCAATAGACCT GCATACG	88	60
Homeobox TF	*AsHomeobox* : c526089_g4	AAATCTTGGTGGGT TTGCTG	TCTGCAGGCTTCAG GTTTTT	97	60
CAM					
Phosphoenolpyruvate carboxylate	*AsPPC* : c489202_g2	TATGCAGACTGAGC GACAGG	CAGAACCAGACTTC CCTTGC	111	60
Phosphoenolpyruvate carboxylase kinase	*AsPPCK* : c477309_g1	TTGAGGAGGATGCT CACTAGG	ACGGGTGCCTCAG GACTT	67	60
Control circadian clock					
GIGANTEA	*AsGI :* c597124_g1_i1	CAAGGGATTGCTTC CATGTT	GCAAATCAACAGCA GACGAA	115	60
TIMING OF CAB EXPRESSION1	*AsTOC1 :* c586374_g1_i7	GGAAGCAGGCATA CACATCA	AACCCACGAGTAGC ATGAGC	124	60
CIRCADIAN CLOCK ASSOCIATED1	*AsCCA1 :* c598520_g2_i2	GCAGGAGCATCTTG GAAAAG	CCTCTTCGGACTGT TTGACC	148	60
Circadian clock associated 1	*AsCCA1 :* c598520_g2_i2	GCAGGAGCATCTTG GAAAAG	CCTCTTCGGACTGT TTGACC	148	65
Cell wall invertase	*As_cwINV :* c589904_g1_i1	GCACAAGTGCAACT GTGAAAA	GCCACTGAATGCGT GTAAAA	142	60
Reference					
Polyubiquitin 10	*AsUBQ10 :* c520612_g2_i3	GCTTCTGAGGGAGT GTCGTG	GGAGCTGGCAAGC AAATGT	108	68

2.4.2 PCR efficiency and melting curve analysis

To determine Q-RT-PCR efficiency for each pair of designed primers, a cDNA pool reverse transcribed from a mix of RNA in an equal amount of all samples was prepared. The cDNA pool was diluted into a series of dilutions: ×1, 1:10, 1:100, 1:1,000, and 1:10,000. Then these serially diluted cDNAs were quantified for the threshold cycle (C_t) values using the Q-RT-PCR techniques according to section 2.4.3 with each pair of primers listed in Table 2.5. The measured C_t values were plotted against the known concentration of each cDNA sample in the dilution series in order to generate a standard curve. The PCR efficiency was calculated from the slope of the standard curve using the following equation: Efficiency = $10^{(-1/\text{slope})} - 1$.

Melting curve (dissociation curve) was plotted using the change in fluorescence detected when double-stranded DNA (dsDNA) with incorporated SYBR Green dye molecules dissociated (melted) into single-stranded DNA (ssDNA) due to the increase of the temperature of the reaction. It was plotted against temperature, and then the $-\Delta F/\Delta T$ (change in fluorescence/change in temperature) was plotted against temperature to illustrate the melting dynamics. The PCR efficiency was automatically calculated and melting curve was automatically plotted using MxPro 4.1 QPCR Software that came with the Mx3005P qPCR System (Agilent Technologies, USA).

2.4.3 Q-RT-PCR Techniques

Total RNA was reverse transcribed to cDNA using the method described in section 2.3.2. 1 µl diluted cDNA was mixed with PCR master mix containing 5 µL SensiFAST™ SYBR® No-ROX Mix 2× (Bioline, UK), 0.4 µL forward and 0.4 µL reverse primer (final concentration 0.4 µM) and 3.2 µL nuclease-free water. Three technical replicates (PCR reactions) were carried out for each biological replicate sample making the total amount of reactions 3 times greater than the

amount of samples. Three biological replicates were used for each sample. These reactions were amplified simultaneously using Mx3005P qPCR System (Agilent Technologies, USA) with the following cycling parameters: 95 °C for 2 min, anneal at a primer specific annealing temperature for 10 s, extend at 72 °C for 10 s, denature at 95 °C for 5 seconds. The annealing, extending and denaturing steps were repeated for 39 cycles making a 40 cycles in total. The extra melting cycle: 95 °C for 1 min, anneal at a primer specific annealing temperature for 30 s, and denature at 95 °C for 30 seconds, was added at the end of the last cycle. Annealing (melting) temperature was adjusted for each primer to obtain the proper % efficiency. The cDNA pool used in section 2.4.2 was added into the same PCR plate along with the experimental samples and set as a calibrator sample.

2.4.4 Transcript abundance quantification

SYBR® Green I dye, a fluorescent DNA-binding dye was used to bind to the double-stranded DNA. The amount of cDNA was measured after each cycle as signal from the SYBR Green I fluorescent dye should increase the fluorescent signal according to the amount of dsDNA PCR product. SYBR Green only fluoresces when intercalated into dsDNA. This generated quantitative information on the initial quantity of transcript (cDNA) in the samples through calculation of threshold cycle (C_t) values. Comparative quantification algorithms—$\Delta\Delta C_t$ was applied to determine transcript level of the gene of interest (Livak and Schmittgen, 2001). C_t values of the gene of interest in experimental sample(s) and calibrator (cDNA pool) were normalized to C_t values of a reference gene (*UBQ10*). The resulting $\Delta\Delta C_t$ value was used to determine the fold difference in expression (Fold difference = $2^{-\Delta\Delta C_t}$). These relative fold-change values of samples were used for the analysis. Quantification process was carried out using MxPro 4.1 QPCR Software that came with the Mx3005P qPCR System (Agilent Technologies, USA).

2.5 Immuno-blot analysis of proteins associated with CAM

2.5.1 Protein extraction and determination

Protein extraction for immunoblotting in *A. sisalana* followed a method developed and optimized for other Agave species by Prof. Anne Borland's PhD student, Dalal Albaijan, University of Newcastle, UK. Approximately 250 mg frozen ground leaf tissue was placed in 500 µl extraction buffer on ice: 280 µl 1M Tris pH 8.3, 100 mM NaCl, 50 µl 1M Dithiothreitol and Protease inhibitors: 10 µl 500mM PMSF(dissolved in DMSO), 40 µl E-64, 40 µl Leupeptin, 40 µl protease inhibitor cocktail (SIGMA (P9599-5ML), and 40 µl 200mM EDTA. Plant tissue was placed on top of the extraction buffer and the tube was left open for 1 minute to allow excess liquid nitrogen to evaporate. The sample was then mixed rapidly with the buffer by inversion and shaking. Tubes containing each extract were then centrifuged for 10 mins at 4˚C and 13,000 g. The supernatant was transferred to a fresh 1.5 mL Eppendorf tube, centrifuged for 10 min at 4˚C and 13,000 g. This step was then repeated once. Glycerol was added to the supernatant to achieve a final concentration of approximately 10 % glycerol (v/v). The extract was snap-frozen in liquid nitrogen and stored at -20˚C until use. The protein content of each extract was determined using Bradford reagent according to the standard procedures (Sigma-Aldrich®, UK).

2.5.2 SDS-PAGE gel electrophoresis

The protocol for sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was carried out according to the standard methods (Laemmli, 1970). The protein extract was mixed 1:1 with 2X sample loading buffer and boiled for 5 min at 100 °C before loading into the gels. The 2X sample loading buffer was prepared as follows; 2.0 ml 0.5 M Tris-HCL pH 6.8, 1.6 ml

100% Glycerol, 3.2 ml 10% SDS, 0.8 ml 2-mercaptoethanol, and 0.4 ml 0.1% (w/v) bromephenol blue in water. Polyacrylamide gels (12% resolving) were prepared as follows:

Resolving gel	**Stacking gel**
3.3 ml H_20	7.7 ml H20
2.5ml 1.5M TRIS-HCl pH 8.8	1ml 0.5M TRIS-HCl pH 6.8
4.2ml Acrylamide (30%)	1.3ml Acrylamide (30%)
100µl 10% SDS	100µl 10% SDS
10µl TEMED	20µl TEMED
100µl AMPS 10% (w/v) *	100µl AMPS 10% (w/v)*

* added just before use.

Different amounts of total protein were loaded into the loading wells of the polyacrylamide gels depending on the abundance of the target protein as follows: 5 µg for abundant PEPC, 10 µg for PPDK and α- and β-NAD-ME, and 20 µg total protein for phosphorylated PEPC, NADP-ME and phosphorylated PPDK. Gels were subjected to electrophoresis at 125V at room temperature using a Mini PROTEAN 3 system (Bio-Rad). Protein marker (PageRuler™ Prestained Protein Ladder, 10 to 180 kDa, Thermo Scientific) was added to the gel. The gels were run until the visible protein marker reached the bottom of the gel. Gels were then stained to visualise protein bands by placing in gel stain buffer consisting of 1.25 g Coomassie Blue 'R' (sometimes called Brilliant Blue), 222 ml Methanol, 225 ml H_2O, and 50 ml acetic acid, slowly and continuously shaken overnight. Gels were then destained in buffer: 1200 ml H_2O, 600 ml ethanol, and 100 ml acetic acid for 4-5 h with a change of destain buffer every 30 min.

2.5.3 Blotting

Protein bands separated on the SDS-PAGE gels and the nitrocellulose membrane to which the proteins were to be transferred were soaked in transfer buffer (no longer than 15min for gels): 25 mM Tris, 192 mM glycine, 0.1% (w/v) SDS, and 10% (v/v) methanol (pH 8.3). The gels were then transferred onto nitrocellulose membrane. The gels and nitrocellulose membrane were put in a sandwich containing materials in following order: fibre pad, filter paper, nitrocellulose

membrane, gel, filter paper, and fibre pad. The sandwich was placed into electrophoretic transfer unit (Mini-PROTEAN 3, Bio-Rad),) with a stir magnet at the bottom for blotting in cold room at 100V for 1 hour or 90V for 1.5-2 h or 30V overnight. The sandwich had to be placed in the correct direction to allow the protein to transfer into the membranes. The blotted nitrocellulose paper was then checked to determine whether the pre-stained markers had transferred by soaking in Ponseau S stain (0.1% (w/v) ponseau S dissolved in 5% acetic acid) until the pre-stained markers were visualised. The blotted membranes were rinsed off with TBS-Tween (50 mM Tris pH 7.5, 150 mM NaCl, and 0.1% Tween 20 (v/v).

2.5.4 Blocking

Membranes were blocked by placing in 5% Marvel dried skimmed milk in TBS-tween on a shaker for 2 h at room temperature or overnight in the cold room. The membranes were then washed with TBS-tween 3 times for 10 min each time on a shaker.

2.5.5 Antibodies

Membranes were given a quick rinse with TBS-tween and then placed in a 50 ml tube containing primary antibody in 1:100 dilution (TBS-tween) for α- and β-NAD-ME and NADP ME, 1:1000 for PPCK (phosphorylated PEPC), 1:3,000 for phosphorylated PPDK, 1:5,000 for PEPC, and 1:10,000 for PPDK. The tubes incubated on a roller mixer for 2 h at room temperature then washed for 3 x 10 min with TBS Tween on a shaker and drained. The primary antibodies (antisera) were kindly provided by the following: PEPC by Hugh G. Nimmo, University of Glasgow (Nimmo *et al.*, 1986), phosphorylated form of PEPC by Cristina Echevarría, Universidad de Sevilla (González *et al.*, 2002; Feria *et al.*, 2008), PPDK and phosphorylated PPDK by Chris J. Chastain, Minnesota State University, Moorhead (Chastain *et al.*, 2000; Chastain *et al.*, 2002), and α- and β NAD-ME by Maria F. Drincovich, Universidad Nacional de

Rosario (Tronconi *et al.*, 2008). The membranes were then placed in the secondary antibody (ECL™ anti-Rabbit IgG, horseradish peroxidase-linked F(ab)2 fragment (from donkey) (Life Technologies™, UK) using a 1:5,000 dilution in TBS-tween for NADP-ME and phosphorylated PPDK, 1:10,000 for PPCK, PPDK, α- and β-NAD-ME, and 1:20,000 for PEPC. The tubes incubated on a roller mixer for 2 h at room temperature then washed for 3 x 10 min with TBS Tween on a shaker and drained.

2.5.6 Detection

An equal volume of each Pierce™ ECL Western Blotting Substrate detection reagent 1 and 2 (Life Technologies™, UK) was mixed just before use in dark room. It was then poured onto the membranes. The membranes were constantly covered by the mixed reagent liquid by pipetting and moving the tray for exactly 1 minute, put in a transparent plastic bag and completely drained. In the dark under the safelight, an X-ray film was placed onto the plastic bags with the membranes inside. The X-ray film was then exposed in a cassette for approximately 1-2 min, developed, fixed and washed. The processed film was hung to dry and digitally scanned for the image analysis and presentation in the results.

2.6 CO_2 exchange analysis using infra-red gas analyser system

The planned gas exchange experiment of the different *A. sisalana* leaf sections to determine where and when dark CO_2 fixation associated with CAM was occurring was not possible before the end of this project due to the delayed process of the repair of the multichannel gas exchange system which was going on for over a year. The representative *A. sisalana* gas exchange results presented in this report were obtained from an experiment performed previously by Susie Boxall (LD data for different sections of the *A. sisalana* leaf) and James Hartwell (LD followed by LL data for the CAM leaf tip of *A. sisalana*), University of Liverpool.

2.7 Enzyme linked spectrophotometric assays for the measurement of soluble sugars in extracts of total soluble metabolites from *A. sisalana* leaves

The method used in this experiment was guided by the method described in "UV method for the determination of sucrose, D-glucose and D-fructose in foodstuffs and other materials", Boehringer Mannheim (R-Biopharm AG, Germany). 100 mg frozen ground sample tissue was continuously homogenised under liquid nitrogen in a mortar using a pestle, and 1 ml 80% (v/v) ethanol was added during the grinding process until the cells were homogenised and lysed completely. The lysate was transferred into a 15 ml tube. Two ml of 80% (v/v) ethanol was added and mixed well. The lysate was extracted by incubating the tube for 1 h at 70˚C in a water bath. During the incubation at 20 min, the tube was centrifuged at 4,500 g for 1 min and the supernatant was transferred into a new 15 ml tube. Next, 2 ml 80% (v/v) ethanol was added to the pellet. The tube was mixed to resuspend the pellet, and returned to the water bath at 70˚C. At 40 and 60 min, the process of centrifugation and supernatant transfer was repeated. After 1 h of incubation, the total of 7 ml of supernatant was collected, transferred into a glass tube and dried using a vacuum concentrator (Savant™ SPD1010 SpeedVac™, Thermo Scientific, UK) for approximately 2h at 60 ˚C or until the lysate was completely dried but not caramelised (clear but not watery nor brown). The vacuum concentrator was set at the following setting: radiant chamber heat (RC): on, vacuum pressure: 5.0-5.1, and manual run mode. The drying process was checked frequently to prevent caramelisation.

The dried extract was dissolved in 100 µl buffer B (10 mM imidazole pH 6.9, 5 mM $MgCl_2$) using a sterile plastic stick with loop to enhance dissolving. The extract was diluted to 0.5 % in Buffer B. 5 µl of extract was added into in microtiter plate wells containing 195 µl sugar assay

cocktail. To assay 100 wells, the sugar assay cocktail contained 97.74% (v/v) Buffer B, 1.02 mM ATP, 2.04 mM NADP, and 0.4 U G6PDH. The solution was incubated at 37°C for approximately 20 min or until the reaction was completely finished. This depended on the concentration of sample or enzyme activity. A kinetic session step on the Multiskan™ programme was set to perform a kinetic measurement in order to be able to determine the reaction of the enzyme. When the line graph of the reaction of the enzyme plateaued and remained at the plateau level, it implied that the reaction was finished. ABS (absorbance) was measured at 340 nm (A1) using a Multiskan™ GO Microplate Spectrophotometer (Thermo Fisher Scientific, USA). This first step of the sugar assays transformed all of the free Glucose-6-Phosphate to Gluconate-6-Phosphate and therefore gave a zero-point ABS reading. 0.3 U HK (Hexokinase) was added to the microtitre plate wells, mixed and incubated at 37 ˚C for 20 min or until reaction finished (the enzyme activity plateaued and remained steady). ABS was measured at 340nm (A2). 0.35 U PGI (phosphoglucose isomerase) was add to the microtitre plate wells, mixed and incubated at 37 ˚C for 20 min or until the reaction finished (the enzyme activity plateaued and remained steady). ABS was measured at 340nm (A3). 0.8 U Invertase (β-Fructosidase) was add to the microtitre plate wells, mixed and incubated at 37 ˚C for 20 min or until the reaction finished (the enzyme activity plateaued and remained steady). ABS was measured at 340nm (A4).

To calculate the content of sugars, the following equation was used:

$c = \frac{\mathrm{V} \cdot \mathrm{MW}}{e \cdot d \cdot v \cdot 1000} \cdot \Delta A\,[g/l]$ where:

c = concentration

V = final volume [ml]

v = sample volume [ml]

MW = molecular weight of the sugar to be assayed [g/mol]

d = light path [e.g. 1cm with K-factor set depends on plate/curvette, machine]

e = absorption coefficient of NADPH at 340nm = 6.3

$\Delta A_{glucose}$ = A2 – A1

$\Delta A_{fructose} = A3 - A2$

$\Delta A_{sucrose} = (A4 - A3)/2$

The unit of sugar content (g/l) was converted to µg/g fresh weight of the leaf tissue used in the original extraction in ethanol.

2.8 Enzyme linked spectrophotometric assays for the determination of malate concentrations in extracts of soluble metabolites from *A. sisalana* leaves

The sample tissue extraction method was the same as described in Section 2.7. The extract was then assayed for malate content using the method from Möllering, (1974). Assay mix for 1x96 well plate was prepared as follows: 80% (v/v) Buffer B, 30 mM Glutamate (pH 10), 2.7 mM NAD, 0.2 U GOT. 189 µl of assay mix was pipetted into each well of a 96-well microtitre plate. 10 µl sample was added into the wells. An initial ABS measurement was made at 340 nm (A1). 1 µl MDH was then added into each well. Plates were incubated at room temperature and the MDH reaction was allowed to process for 10 min or until reaction finished (the enzyme activity plateaued and remained steady). ABS was measured at 340nm (A2). Malate content was calculated as follows:

$c = [(\Delta A/E)*0.0002*10*d]/g$

Where:

c = malate content (mmol g malate/g fresh weight)

ΔA (A2-A1) = difference of ABS after adding MDH

E = extinction coefficient (2.9107)

0.0002 (L) = 200 µl of total assay

10 = multiplied by 10 to get mmol g malate in 100 µl sample extract

d = dilution factor (e.g. 100 if diluted to 1:100)

g = fresh weight (g)

The concentration of malate was then converted into µmol g fresh weight by simply multiplying by 1000.

2.9 Ilumina Hi-Seq RNA-sequencing

Due to the high cost of sequencing library preparation from multiple replicated total RNA samples and the subsequent high cost of the Illumina sequencing for each library, only a subset of leaf segments from biological triplicate samples collected at the 10:00 light (2 h before dusk) and 22:00 dark (2 h before dawn) time points were chosen for RNA-seq anaylsis using the Illumina Hi-Seq system. The following leaf samples were chosen: white basal, pale green basal, and dark green tip part of the youngest fully expanded leaf (Figure 2.2, leaf segments 1, 2 and 3 respectively) sampled at 10 h (light) and 22 h (dark) using 3 biological replicates at each time point, leading to a total of 18 RNA samples (see section 2.1.2). The selection method of leaf segments and time points was based on the results of preliminary RT-PCR based transcript level measurement, described in chapter 3.

2.9.1 DNase treatment and quality control of RNA

The RNA isolated from the 18 samples was DNase-treated in solution using TURBO™ DNase Treatment and Removal Reagents (TURBO DNA-free™, Ambion® by Life Technologies™) according to the manufacturer's instructions. The DNase treated RNAs were quantified using Qubit Fluorometer (Qubit® fluorometer, Invitrogen™ by Life Technologies™). Quality control was performed using the Agilent 2100 Bioanalyzer machine with the RNA 6000 Pico Chip Kit (Agilent 2100 Bioanalyzer, Agilent Technologies Inc.), according to the manufacturer's instructions.

2.9.2 Library preparation, RNA-sequencing and quality control

The DNase-treated RNA samples were submitted to the Centre for Genomic Research (CGR), University of Liverpool for RNA-sequencing using the Illumina Hi-Seq 2500 sequencer and reads quality control. Illumina cDNA sequencing libraries were created using the ScriptSeq™ Complete Kit (Plant Leaf) (Epicentre®, an Illumina company) which included an rRNA depletion step, according to the manufacturer's instructions. The sequenced reads were trimmed for quality control purposes. The raw Fastq files containing reads were trimmed for the presence of Illumina adapter sequences using Cutadapt version 1.2.1 (Martin, 2011). The option -O 3 was used, so the 3' end of any reads which matched the adapter sequence for 3 bp or more were trimmed. The reads were further trimmed using Sickle version 1.200 (Joshi NA, 2011) with a minimum window quality score of 20. Reads shorter than 10 bp after trimming were removed. Statistics were generated using fastq-stats from EAUtils (Aronesty, 2013).

2.10 Comprehensive analysis of RNA-seq data

All bioinformatics works were operated using Linux computer servers via a command-line interface with a number of processing units in order to handle large RNA-seq data processing. This was provided by Centre for Genomic Research (CGR), University of Liverpool.

2.10.1 De novo assembly

De novo assembly was necessary as there was no *A. sisalana* complete genome sequence available during this project. The trimmed Illumina paired-end RNA-seq reads for all 18 samples were assembled using Trinity (Grabherr *et al.*, 2011). The trinity algorithm included 3 different components of software: Inchworm, Chrysalis, and Butterfly, applied in sequential steps. Inchworm assembled the sequenced reads into the unique sequences of transcripts and reported the unique portions of alternatively spliced transcripts. The Inchworm contigs were grouped into clusters and constructed to form the complete de Bruijn graphs for each cluster using Chrysalis. Chrysalis then divided the full read set among these separate graphs. The individual graphs were then processed in parallel using Butterfly. The paths that reads and pairs of reads took within the graph were traced. Butterfly then finally reported full-length transcripts for alternatively spliced isoforms, and excluded transcripts that corresponds to paralogous genes (Grabherr *et al.*, 2011). Option "--SS_lib_type FR" was used as Illumina sequenced reads were strand-specific. The Trinity assembly was evaluated using the Trinity built-in tool (TrinityStats.pl), Quast 3.0 (Gurevich *et al.*, 2013), and Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra *et al.*, 2007).

2.10.2 Annotation

TransDecoder, included in the Trinity package, was used to generate the most likely longest-ORF peptide candidates extracted from the Trinity assembled contigs. Shorter ORFs that overlapped with the longer ones were excluded (Haas *et al.*, 2013). Transdecoder-predicted protein coding regions were used for sequence homologies search using BLASTP in annotation step using Trinotate, an annotation tool included in the Trinity package (Grabherr *et al.*, 2011). Trinotae integrated the different well known functional annotation methods including homology search on known databases (BLAST+/SwissProt/Uniref90), protein domain identification (HMMER/PFAM), protein signal peptide and transmembrane domain prediction (signalP/tmHMM), and comparison to active annotation databases (EMBL Uniprot eggNOG/GO Pathways databases). The data generated using these methods were integrated into SQLite database allowing a fast and efficient search for annotation information. An annotation report containing annotation information from all databases for individual Trinity contigs of the whole transcriptome was generated.

2.10.3 Differential expression analysis

The RNA-seq raw reads were aligned to the *de novo* assembly, previously created using Trinity section, using Trinity built-in tool, RSEM (Li and Dewey, 2011). In the RSEM setting, Bowtie 2, an ultrafast and memory-efficient tool for aligning sequencing reads, was used due to its capability to align longer reads generated from the next-generation sequencing technology such as Illumina (Langmead and Salzberg, 2012). Mapped read counts per Trinity "gene" and isoform per sample as well as TPM and FPKM values were generated. Mapping quality was evaluated using "samtools flagstat" on the alignment output (bowtie2.bam) file. Read counts generated using RSEM, annotation, design, contrast, and FPKM table were created. These

tables were then used as inputs for the differential expression analysis which was performed using edgeR (Robinson *et al.*, 2010) by Dr. Yongxiang Fang, Centre for Genomic Research (CGR), University of Liverpool.

Several contrasts were used for the differential expression analysis with edgeR, described in Table 2.6. The count per million (CPM) value was calculated and used by edgeR to reflect expression levels. In the edgeR analysis, tagwise dispersion values of individual transcripts (Trinity 'genes') were estimated using count data from each sample generated previously using RSEM to enable the estimation of biological variability amongst the samples. The tagwise dispersion values were normalized and used to fit the negative binomial model. Count data was modelled within edgeR using an over-dispersed, Poisson model. An empirical Bayes method was used to moderate the degree of over-dispersion values across transcripts, squeezing the dispersions towards a consensus value. Differential expression of transcripts was then assessed and statistics were calculated (Robinson *et al.*, 2010). Using these methods it was possible to identify the differentially expressed (DE) Trinity 'genes'.

Table 2.6 Contrasts used for the differential expression analysis with edgeR

Contrast	Sample 1 (segment, time)	Sample 2 (segment, time)
1	leaf tip 10:00 L	leaf base 10:00 L
2	leaf tip 10:00 L	leaf base white 10:00 L
3	leaf base 10:00 L	leaf base white 10:00 L
4	leaf tip 22:00 D	leaf base 22:00 D
5	leaf tip 22:00 D	leaf base white 22:00 D
6	leaf base 22:00 D	leaf base white 22:00D
7	leaf tip 10:00 L	leaf tip 22:00 D
8	leaf base 10:00 L	leaf base 22:00 D
9	leaf base white 10:00 L	leaf base white 22:00 D
10	total leaf tip 10:00 L + 22:00 D	total leaf base 10:00 L + 22:00 D
11	total leaf tip 10:00 L + 22:00 D	total leaf base white 10:00 L + 22:00 D
12	total leaf base 10:00 L + 22:00 D	total leaf base white 10:00 L + 22:00 L

L = Light, D = Dark

2.10.4 Identification of novel differentially expressed genes with potential functions in the light/ dark coordination and optimisation of CAM in A. sisalana

From the list of DE genes generated using edgeR, the genes were clustered based on their expression behaviour and presented in a heatmap to aid with the visualisation of the expression patterns of each gene cluster. Venny diagram 2.0.2 (Oliveros, 2007-2015) was also used to group the DE genes expressed in different leaf segments and time points. The DE genes were then sorted in Microsoft Excel based on their transcript abundance values (LogFC), ranked from the highest transcript level in certain leaf segments and time points (FDR < 0.05). The most frequent functionally annotated genes amongst the list of DE genes were analysed using an R script in order to count the frequency of genes with a certain Pfam annotation that occurred in the set of DE genes. Those genes that were annotated as transcription factors (TFs) and/ or DNA-binding proteins, which also had a similar pattern of transcript abundance regulation to that of known CAM genes, were identified using the annotation information of each Trinity "gene" derived previously through Trinotate. The FPKM values of novel genes were plotted to clearly demonstrate the expression (transcript) level in different leaf segments and time points. From this analysis of the list of DE genes, novel genes encoding TFs were selected for follow-up Q-RT-PCR analysis using the complete 24 h LD time course. This allowed not only for validation/ corroboration of the RNA-seq data, but also revealed a more detailed view of the regulation of each discovered TF over the complete light/ dark cycle. This was important as it provided more detailed information about the timing of the peak and trough of the daily oscillation of the transcript abundance of each TF.

Chapter 3

An investigation of the timing and localization of peak transcript abundance for a range of known CAM-associated genes in *A. sisalana*

3.1 Introduction

A number of previous studies in obligate CAM and C_4 species have characterised a clear developmental transition from C_3 to CAM or C_4, either by comparing molecular, biochemical and/ or physiological characteristics in different ages of a dicot CAM leaves (Jones, 1975), or by comparing diagnostic characteristics for CAM or C_4 along the length of monocot leaves (Li *et al*., 2010; Gross *et al*., 2013). For example, Jones, (1975) demonstrated that full CAM was present in the older leaves of the obligate, dicot CAM species *Bryophyllum* (*Kalanchoë*) *fedtschenkoi,* while the young leaves were performing C_3. In more recent molecular work on the same species, the known CAM gene *PPCK* was shown to be under clear developmental control, achieving its full CAM-associated nocturnal transcript peak in the 4^{th} to 5^{th} leaf pair down the stem from the shoot apex and in older leaves (Hartwell *et al.*, 1999).

In the C_4 monocot species maize (*Zea mays*), Li *et al.,* (2010) performed a study of photosynthetic development using the *de novo* transcriptome sequencing approach by measuring changes in gene transcript levels along the length of young leaves, which undergo photosynthetic differentiation along the proximal-distal axis. They found that 64 % of genes were differentially expressed along the length of the leaf that was divided into 4

developmental sections (Li *et al.,* 2010). In the areas of sink-to-source transition tissue, there was an increase in the level of transcripts that were related to mechanisms of photosynthetic development. In the distal areas of the leaf, the cellular processes were found to be nearly exclusively associated with photosynthesis. Furthermore, they demonstrated that the C_4 photosynthetic genes increased from leaf base towards leaf tip (Li *et al.,* 2010). A more recent trancriptomic or RNA-seq study in developing maize studied the gene transcript level changes in relation to germination and early leaf development in maize and revealed that photosynthesis-related transcription factor genes are highly differentially expressed among different stages of the early development of embryonic leaves (Yu *et al.*, 2015).

In *A. sisalana*, a preliminary study of the developmental control of CAM CO_2 fixation and the developmental induction of CAM-related genes including *AsPPC, AsPPCK* and *AsPPDK* was previously undertaken in the Hartwell lab (Boxall, Waller and Hartwell, unpublished). The preliminary results showed that CAM was fully developed in the mature green tissue at the leaf tip. CAM decreased towards the base of the leaf, which is the least mature area in monocot leaves, such that the base of the leaf performed C_3 photosynthesis, fixing atmospheric CO_2 solely in the light period.

The initial scoping experiments described in this chapter were carried out to investigate the timing and localization of peak transcript abundance for a range of known CAM-associated genes within developing leaves of *A. sisalana*. Semi-quantitative RT-PCR was applied to study the transcript abundance level for a range of known positive control CAM, circadian clock and sucrose and fructan metabolism associated genes. The transcript abundance level data presented in this chapter were calculated from experiments involving a single biological replicate of each time point/ leaf section, with three technical replicates used to generate the displayed standard error bars. This lack of biological replicates was an unavoidable

consequence of the limited supply of *A. sisalana* plants at the beginning of this PhD. Luckily, for later experiments in subsequent chapters, a supply of *A. sisalana* adventitious bulbils from flowering plants at a nursery in Portugal was sourced, and so large population of developmentally synchronised, clonal young *A. sisalana* plants could be established in the greenhouses at the University of Liverpool allowing the subsequent chapters to include full biological triplicates. The scoping experiments described in this chapter were however extremely valuable in terms of guiding the selection of the leaf sections and time points that were sampled for the fully biological replicated RNA-seq experiment described in chapter 5. It should be noted that not all of the genes for which PCR primers were designed, and for which scoping RT-PCR experiments were performed, are presented in this chapter, instead only a representative selection of genes which displayed informative results are presented.

3.2 Results and discussion

3.2.1 Initial scoping experiment

An initial experiment was carried out to determine the tissue specificity of the transcripts of a selection of CAM-associated genes (Table 2.4). The first experiment aimed to screen for which part of the whole *A. sisalana* plant expressed the highest level of CAM induction, using a measurement of transcript levels of well-defined CAM genes in different parts of the plant. The experiment was carried out using a 1-year-old plant propagated from side shoots, and grown in a greenhouse under a 16 h light/ 8 h dark cycles. The whole plant was sampled at 6 h after the light started, dissected into the various following sections: meristem, primordial leaves around the meristem, 4 cm long base and tip of youngest leaves above meristem, immature leaves (3rd leaf from core) and mature leaves (8th leaf from the outer-most leaf of the rosette), 2-3 mm thick of fibrous tissue below meristem and above root, fibrous tissue with

root initial attached, rhizome (swollen root), plus the nodes and inter-nodal sections from the stolon left from the mother plant, and stolons detached from a maturing three-year-old plant, making total 15 samples (see Section 2.1.1). The semi-quantitative RT-PCR results for the first preliminary experiment showed that transcript abundance of *AsPPC, AsPPCK*, and *AsPPDK* was quite different through different parts of the plant (Figure 3.1). *AsPPC* showed the highest transcript abundance in the tip of immature leaves (3rd leaf from meristematic cone; sample 6) followed by a slightly lower transcript level in the tip of mature leaves (8th leaf from the outer-most leaf of the rosette; sample 8). The *AsPPC* transcript levels in these two tip sections were much higher than the other parts of the plant (Figure 3.1A). The base of mature leaf (8th leaf from the outer-most leaf of the rosette; sample 7) also exhibited a relatively high level of *AsPPC* transcript relative to the rest of the samples, including meristem (sample 1), leaf primordial from around meristem (sample 2), base and tip of youngest leaves separated from the meristematic cone (samples 3 and 4), base of immature leaf (sample 5), and underground tissues (samples 9-15) (Figure 3.1A). *AsPPDK* also showed a similar high level of transcript abundance in the tip of immature leaves (sample 6; 3rd leaf from meristematic cone) among other samples (Figure 3.1B). The tip of the youngest leaves separated from the meristematic cone (sample 4), the base of immature leaves (3rd leaf from meristematic cone; sample 5), and the base and tip of mature leaves (8th leaf from the outer-most leaf of the rosette; samples 7 and 8) also demonstrated a reasonably high level of transcript relative to other parts of the plants especially meristem (sample 1) and underground tissues (samples 9-15) (Figure 3.1B). Like *AsPPC*, the transcript levels of *AsPPDK* in leaf primordial from around meristem (sample 2) and the base of the youngest leaves separated from the meristematic cone (sample 3) were also relatively low (Figure 3.1B). The high level of *AsPPDK* transcript in the tip of the youngest leaves separated from the meristematic cone (sample 4), immature and mature leaf related tissues (samples 5-8), especially in leaf tip (samples 6 and 8) (Figure 3.1B), and high level of

AsPPC transcript in samples 6-8 (Figure 3.1A), relative to the other parts of the plants might indicate that these CAM genes are most actively transcribed (or their transcripts are most stable) in tissues where source leaf photosynthesis is occurring (Borland *et al.*, 2009). Unlike *AsPPC* and *AsPPDK*, the *AsPPCK* transcript levels in the underground tissues (samples 10-15) were very similar to the levels in the meristem related (samples 1 and 2) and leaf related samples (samples 3-8), except for the fibrous tissues from the zone below the meristem and above the root (sample 9) which showed much lower level of *AsPPCK* transcript abundance compared to the rest of the samples (Figure 3.1C). However, *AsPPCK* transcript in the tip of immature leaves (3rd leaf from meristematic cone; sample 6) still exhibited the highest level when compared to the rest of the samples. *AsPPCK* has been shown in other CAM species to be most abundant and active in the dark period and very low in the light period (Boxall *et al.*, 2005; Hartwell *et al.*, 1999; Taybi *et al.*, 2000). All of the tissues and organs used in this multi-organs/tissues experiment were sampled only in the light period (at 6 h after the light started). This may explain why the transcript abundance of *AsPPCK* was similar in the majority of the tissues investigated here, as the leaf tissues were all sampled in the light when *AsPPCK* might be expected to reach its daily trough in transcript abundance if it is regulated in the same way in *A. sisalana* as reported previously for other CAM species.

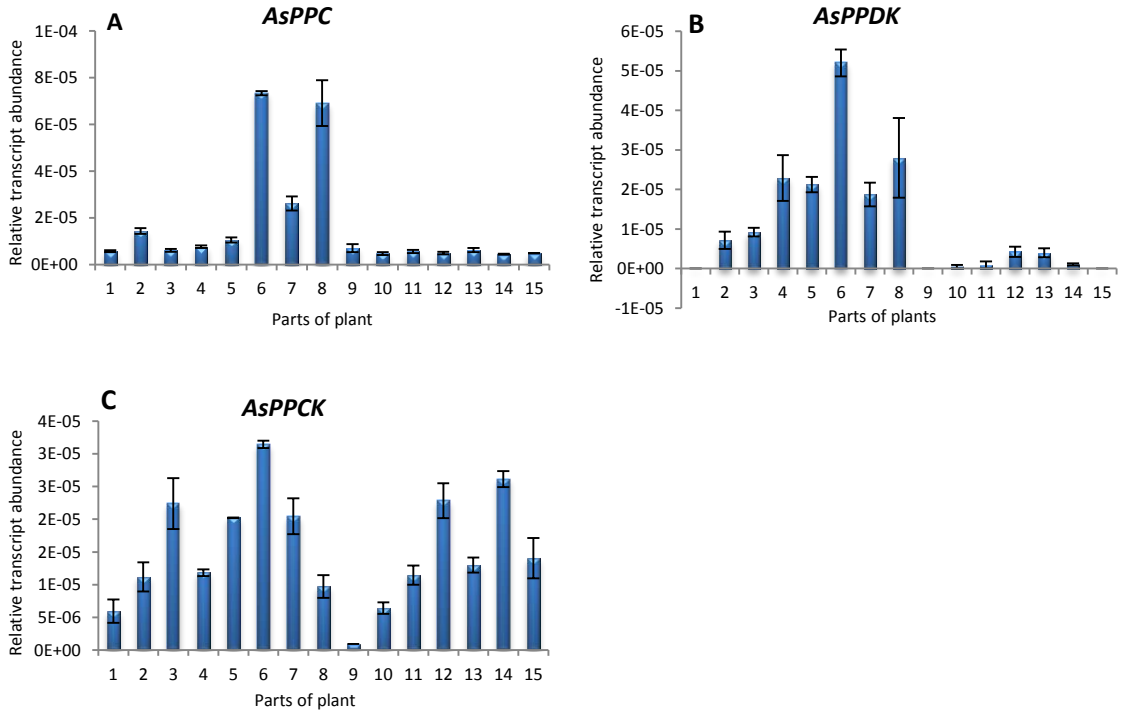


Figure 3.1 Relative transcript abundance level of CAM genes *AsPPC, AsPPDK* and *AsPPCK* in different tissues and organs of *A. sisalana.*

Semi-quantitative RT-PCR was used to determine the relative transcript abundance of CAM genes *AsPPC, AsPPCK* and *AsPPDK* in different tissues and organs of *A. sisalana* sampled at 6 h after the light started. (1) meristem, (2) leaf primordial from around the meristem, 4 cm long (3) base and (4) tip of youngest leaves separated from the meristematic cone, 4 cm long (5) base and (6) tip of immature leaves (3rd leaf from meristematic cone), 4 cm long (7) base and (8) tip of mature leaves (8th leaf from the outer-most leaf of the rosette), (9) 2-3 mm thick cross sections of fibrous tissue from the zone below the meristem and above the root, (10) fibrous tissue with root initials attached, (11) stolon left from mother plant, (12) 4cm long base of stolon taken from a three-year-old plant, (13) 2-3 cm thick samples of nodes from the stolon left from mother plant, (14) rhizome (swollen root), and (15) 4 cm long tip of stolon taken from a three-year-old plant. Transcript abundance values were normalized to the abundance of UBQ10 transcripts amplified from the same cDNA samples.

In the second scoping experiment, approximately 3-year-old plants were sampled in the light (2 h before dusk) and dark (2 h before dawn) period by collecting the youngest fully expanded leaf and measuring the transcript abundance of *AsPPC*, *AsPPCK* and *AsPPDK* using semi-quantitative RT-PCR. *AsPPC* and *AsPPDK* showed similar patterns of transcript abundance. They were both higher in leaf tip in the light compared to the leaf base (Figure 3.2A and B). The difference was most pronounced for *AsPPC* (Figure 3.2A). *AsPPCK* transcript levels were

higher in the leaf base compared to the leaf tip and higher in the dark than light period (Figure 3.2C).

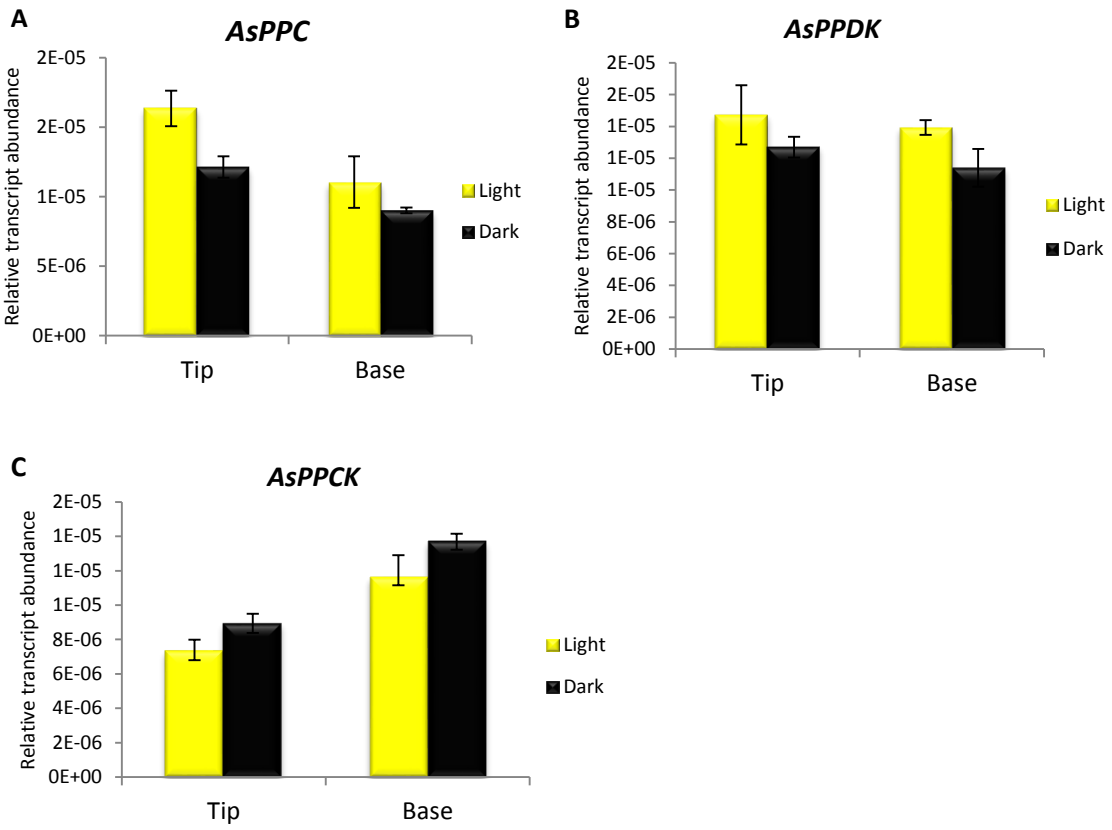


Figure 3.2 Relative transcript abundance level of *AsPPC, AsPPDK* and *AsPPCK* in the leaf tip and base of the youngest fully expanded leaf.

A 2-cm-long section from the youngest fully expanded leaf base and a 5-cm-long section of the leaf tip were sampled from 3-year-old *A. sisalana* plants at 2 h before dusk (light) and 2 h before dawn (dark) in 12:12 light/ dark cycles. Transcript abundance was determined using semi-quantitative RT-PCR; values were normalized to the reference gene UBQ10.

In fully mature leaves of the same approximately 3-year-old plants, again *AsPPC* and *AsPPDK* transcripts showed similar pattern to each other (Figure 3.3A and B), but a different pattern compared to the result for the youngest fully expanded leaves (Figure 3.2A and B). Here, they were slightly higher in leaf base than leaf tip, but still higher in the light than dark period. *AsPPCK* also showed higher transcript level in leaf base compared to leaf tip in the light period (Figure 3.3C).

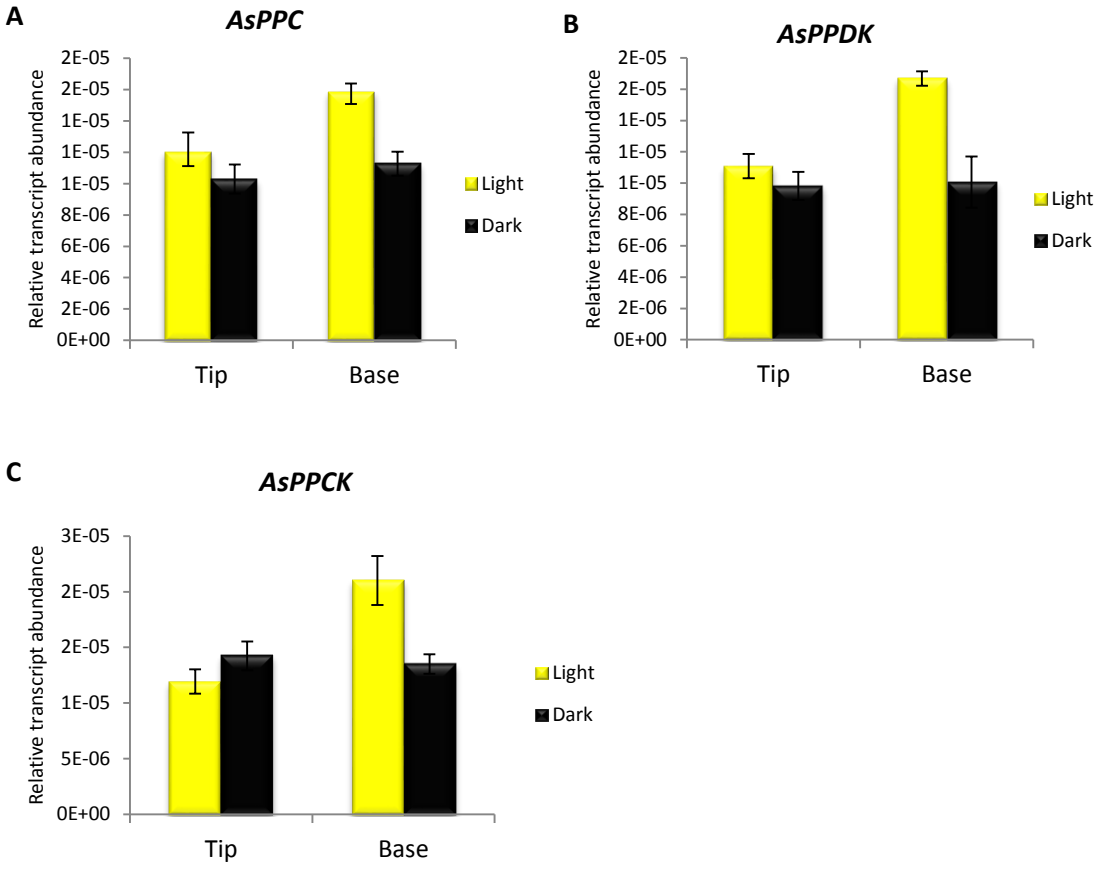


Figure 3.3 Relative transcript abundance level of *AsPPC*, *AsPPDK* and *AsPPCK* in the leaf tip and base of mature leaf.

A 2-cm-long section from the mature leaf Base and a 5-cm-long section of the leaf tip were sampled from 3-year-old *A. sisalana* plants at 2 h before dusk (light) and 2 h before dawn (dark) in 12:12 light/ dark cycles. Transcript abundance was determined using semi-quantitative RT-PCR; values were normalized to the reference gene UBQ10.

Noticeably, in the mature leaf experiment, *AsPPC* and *AsPPDK* transcript levels were higher in the leaf base in the light compared to the leaf tip in the light, although the transcript levels were similar between the tip and base in the dark (Figure 3.3A and B). This was in contrast to the result for the young leaves from the same plants where *AsPPC* and *AsPPDK* were higher in leaf tip than leaf base (Figure 3.2A and B). This might be because CAM genes were already functioning in the base of mature leaves. Based on these preliminary results from scoping experiments, it was decided to investigate the developmental control of CAM genes in the youngest leaves from younger *A. sisalana* plants (approximately 1.5-years-old) (Figure 3.4).

In this experiment, *AsPPC* showed higher transcript abundance in the leaf tip than leaf base in the light period whereas levels were very similar in leaf tip and base in the dark (Figure 3.4A). *AsPPDK* was also more abundant in the leaf tip than the leaf base in the light. Like *AsPPC*, *AsPPDK* was very similar in leaf tip and base in the dark (Figure 3.4B). In terms of light/ dark regulation, *AsPPC* in the leaf tip was higher in the light than in the dark samples. This was in contrast with the leaf base where it was higher in the dark than light samples (Figure 3.4A). *AsPPDK* in the leaf tip was not distinguishable between the light and dark samples, considering also the large error bar for the leaf tip light sample that overlapped with the dark sample (Figure 3.4B). However, in the leaf base, *AsPPDK* was higher in the dark than in the light samples. However, *AsPPCK* showed higher transcript level in leaf base compared to leaf tip in the dark period (Figure 3.4C). When comparing transcript levels between the light and dark periods, all genes showed similar patterns. In leaf tip, they were all higher in the light period. In the leaf base, they were higher in the dark period (Figure 3.4).

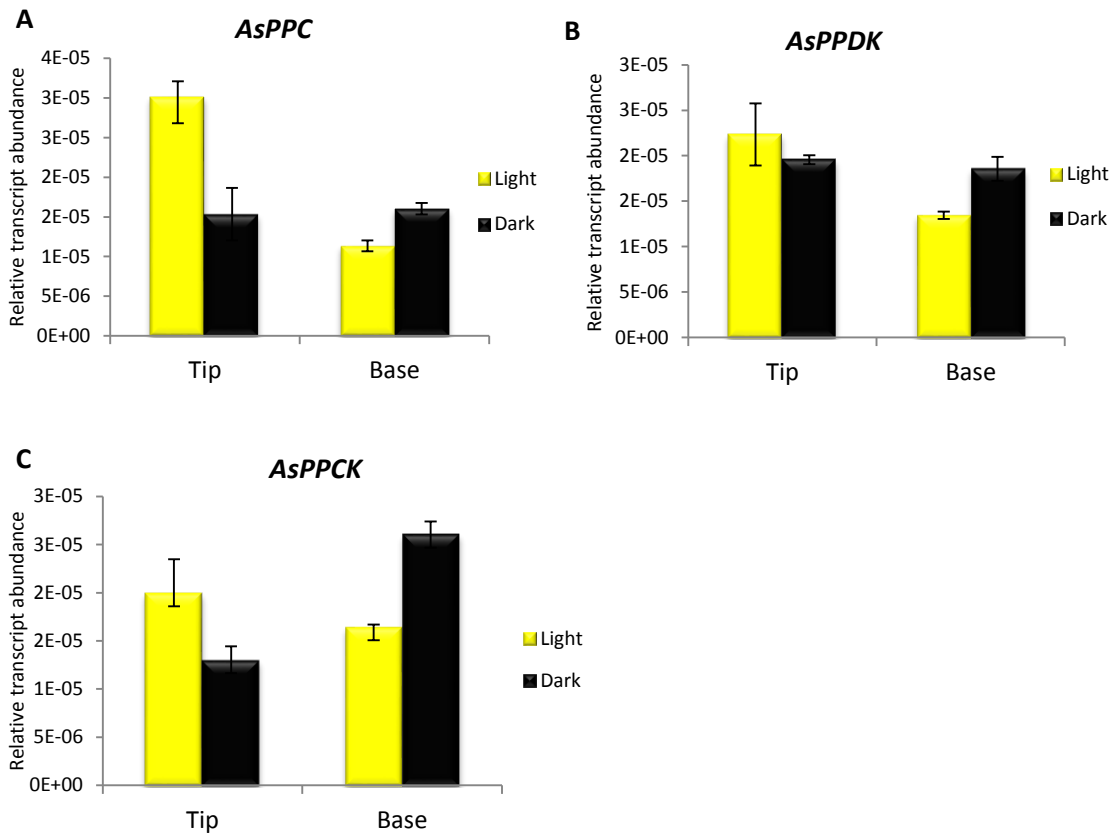


Figure 3.4 Relative transcript abundance level of *AsPPC*, *AsPPDK* and *AsPPCK* in tip and base of youngest fully expanded leaf.

A 2-cm-long section from the youngest fully expanded leaf base and a 5-cm-long section of the leaf tip were sampled from 1.5-year-old *A. sisalana* plants at 2 h before dusk (light) and 2 h before dawn (dark) from plant grown under 12:12 light/dark cycles in a Snijders Microclima MC-1000 growth cabinet. Transcript abundance was determined using semi-quantitative RT-PCR; values were normalized to reference gene UBQ10 transcript levels.

Due to the fact that the sampling times in the above experiments were at only 2 h before dusk (light) and 2 h before dawn (dark), the variation of each gene over the complete 24 h light/dark cycle was determined. A further experiment was undertaken to address this aspect of the regulation of *AsPPC*, *AsPPCK* and *AsPPDK* in *A.sisalana*. Leaves were sampled from 1.5-year-old plants at 4 h intervals over the 12:12 light/dark cycle in order to investigate in greater detail the timing of peak transcript abundance for the previously tested CAM genes plus a selection of circadian clock and sugar metabolism associated genes. In addition to the three CAM marker genes studied in the previous experiments, results are presented for genes that

showed the most pronounced and reproducible differences in terms of their differential regulation between different sections of the leaf, and over the light/ dark cycle. *AsPPC* and *AsPPDK* transcript abundance was higher in the leaf tip than the leaf base throughout the light/ dark time course (Figure 3.5A and C). In contrast, the transcript level of *AsPPCK* peaked in the leaf base at 10:00 (end of light), and stayed relatively high during the dark period (Figure 3.5B). In the leaf tip, *AsPPCK* transcript levels peaked 4 h later, at 14:00 (2 h into the dark), than in the leaf base (10:00; 2 h before the end of the light; Figure 3.5B). *AsPPDK* transcript levels increased throughout the light period, peaking at the beginning of the dark period and declining throughout the remainder of the dark period in both the leaf base and the leaf tip, but the abundance was consistently higher in the tip region throughout the light/ dark cycle (Figure 3.5C). This pattern was similar to the light/dark regulation of the *AsPPC* transcript level, although *AsPPC* levels peaked 4 h earlier in the base (10:00 light) than the tip (14:00 dark) revealing a phase delay in the timing of the daily peak in *AsPPC* transcript levels in the leaf tip relative to the leaf base (Figure 3.5A). This delay in the timing of the *AsPPC* peak in the leaf tip correlated well with the delay in the *AsPPCK* peak in the tip relative to the base (Figure 3.5A and B).

In addition to the CAM genes, the investigation of the transcript levels of clock and sugar metabolism genes also revealed some interesting results. As a representative of the gene within the central circadian clock mechanism, *GIGANTEA* (*GI*) was used. This gene is believed to function as part of the evening loop of the central oscillator in the current model of the central circadian clock in the model plant *A. thaliana*, and was therefore expected to peak in the evening in *A. sisalana*. *AsGI* transcript levels were low at the beginning of the light period, increased throughout the light period, peaked at the end of the light in the leaf base or beginning of the dark in the leaf tip, and declined throughout the dark period (Figure 3.5D). It is interesting to note that *AsGI* peaked 4 h earlier in the leaf base than in the leaf tip, in a

manner that correlated well with the phase delayed peak of *AsPPC* and *AsPPCK* in the leaf tip relative to the leaf base (Figure 3.5A and B).

Although, a number of sucrose and fructan metabolism associated genes were screened using the cDNA samples from this leaf base and leaf tip experiment using 1.5-year-old plants, data is only presented for a putative *CELL WALL INVERTASE* (*As_cwINV)* gene. *As_cwINV* transcript levels were very low to undetectable at all time points in the leaf base. In the leaf tip, *As_cwINV* levels were low throughout the light period, increased dramatically at the beginning of the dark period, and peaked at both the middle and the end of the dark period (Figure 3.5E). If these changes in transcript levels results in altered levels of the encoded protein and correlate with its activate, then these findings suggest that either sucrose and/ or fructan turnover was activated in the leaf tip specifically in the latter half of the dark period, whereas this sucrose or fructan turnover activity was not required in the leaf base of the same leaf.

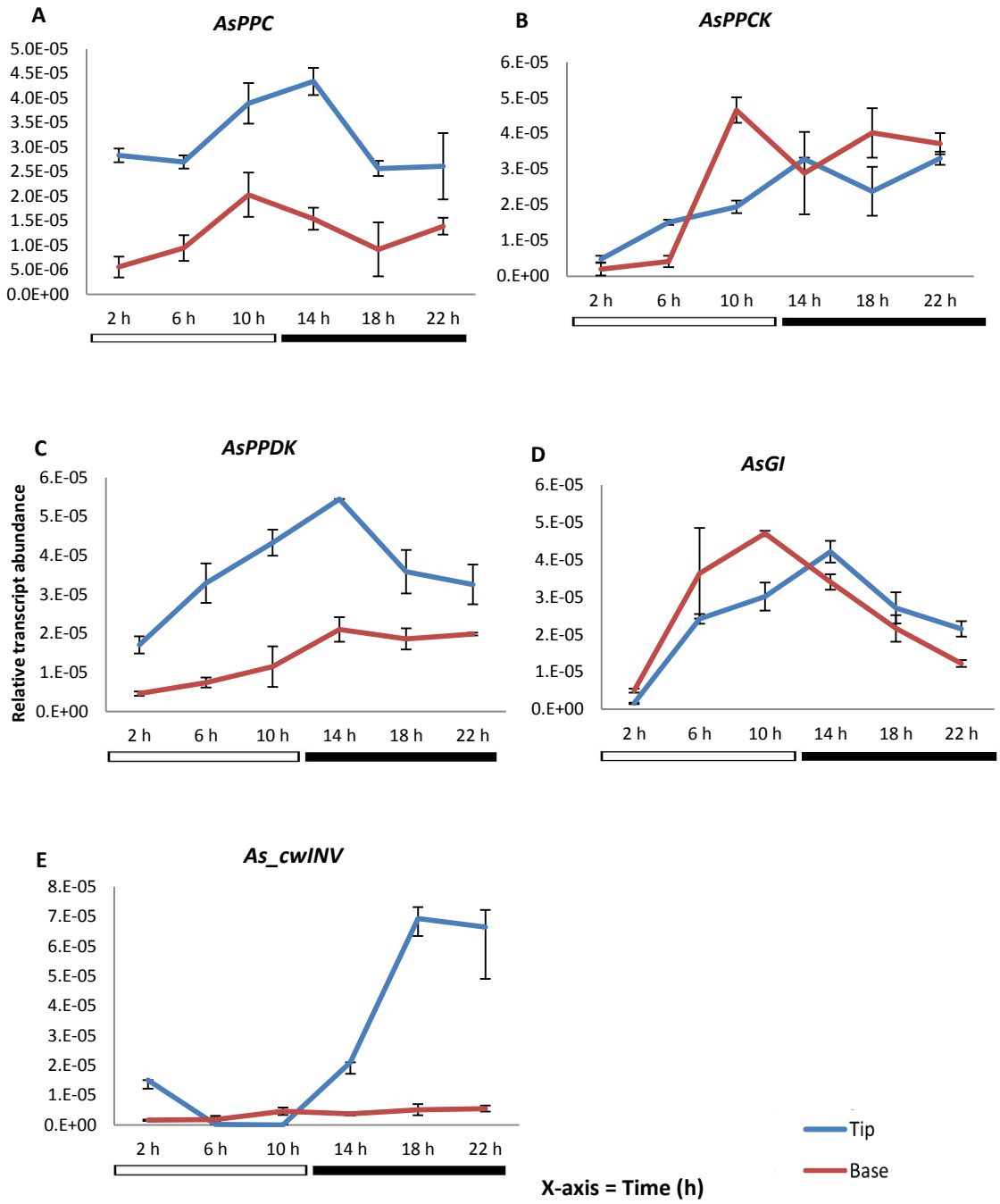


Figure 3.5 Relative transcript abundance level of CAM genes including *AsPPC* (A), *AsPPCK* (B) and *AsPPDK* (C), circadian clock *AsGI* (D), and sucrose related gene *As_cwINV* (E) in tip and base of youngest fully expanded leaf.

A 2-cm-long section of the leaf base of the youngest fully expanded leaf and a 5-cm-long section of the leaf tip, were sampled from 1.5-year-old *A. sisalana* plants sampled at 4 h intervals throughout a 12:12 light/dark cycle. Transcript abundance was determined using semi-quantitative RT-PCR; values were normalized to abundance of UBQ10 transcripts.

The overall outcome of these initial scoping experiments was that the results obtained for the CAM genes *AsPPC* and *AsPPDK* (Figure 3.5A and C) were consistent with those obtained in a previous study by Waller, Boxall and Hartwell (unpublished). *AsPPC* and *AsPPDK* transcript levels were found to be higher in the leaf tip relative to the leaf base, which correlated with higher levels of CAM photosynthesis measured previously in the leaf tip using an infrared gas analyser system. The transcript differences between leaf base and leaf tip for *AsPPC* and *AsPPDK* were most obvious in the light period and these results were consistent with previous finding in divergent CAM species including *M. crystallinum* and *K. fedtschenkoi* (Cushman *et al.*, 2008; Dever *et al.*, 2015). The leaf tip of *A. sisalana* was demonstrated previously to be performing full CAM using gas exchange analysis with an infra-red gas analyser, while the leaf base performed C_3 and developed CAM photosynthesis as the leaf developed from the base towards full photosynthetic competence at the tip (Boxall and Hartwell, unpublished results). The *AsPPC* and *AsPPDK* results obtained were most consistent with the previous data generated in the Hartwell lab. *AsPPCK* was expected to peak in the dark period in CAM leaf tissues according to the previous studies undertaken in the Hartwell lab (Waller, Boxall and Hartwell, unpublished). In the experiments reported here, *AsPPCK* did not show transcript abundance steadily higher in the dark than light period, although it showed higher transcript level in the dark in the experiment on young leaf of mature plant (Figure 3.2), and in the leaf tip sampled every 4 h over the light/ dark cycle (Figure 3.5B). *AsPPCK* also showed different pattern of regulation in the tip-base comparison relative to the other CAM genes studied here, namely that it was not always found to be higher in leaf tip relative to the leaf base. The transcript abundance of *AsGI* peaked either at the end of light in the leaf base, or beginning of dark in the leaf tip (Figure 3.5D). This result was consistent with previous studies in *Arabidopsis* which found that *AsGI* transcript abundance peaks around 8-10 h after dawn (Park, 1999;

Fowler *et al.*, 1999). The previous work in *A. sisalana* by the Hartwell lab (unpublished) also found a similar result.

The putative sucrose degradation gene *As_cwINV* showed an interesting and very clear result, namely its transcript level in leaf base was very low to undetectable at some time points while it increased dramatically and peaked in the middle of the dark and at the end of the dark period in the leaf tip (Figure 3.5E). This result is very well consistent with the findings that sucrose is believed to be the storage carbohydrate for glycolysis to supply the PEP for nocturnal CO_2 fixation by PEPC in Agave species (Christopher and Holtum, 1996; Wang and Nobel, 1998). Thus, *As_cwINV* may function in the leaf tip to breakdown sucrose to release fructose and glucose, which could in turn act as a substrate for the generation of PEP as substrate for PEPC in the dark.

It is important to note that for the 4 h-interval time course experiment on young plants (1.5-years-old; Figure 3.5) the sample collected at 2 h into the light period was from a larger plant than the plants sampled for the other 5 time points. It had a longest leaf of 54 cm, while the other 5 plants had similar length of leaves that were approximately 40 cm long. It would have been better to use 6 plants of the same size, but due to a lack of suitable plants at the time, the experiment had to be carried out this way.

3.2.2 Scoping experiment using the total RNA samples that were subsequently used for the main Illumina RNA-seq experiment

The preceding scoping experiments demonstrated that the strongest differential regulation of CAM genes between leaf base and tip was found in young plants and in the youngest fully expanded leaf (Figure 3.2, 3.4 and 3.5). Thus, these results directed the subsequent experiments to move on to investigate of the developmental control of CAM genes in the

youngest fully expanded leaves of younger *A. sisalana* plants. As the supply of young *A. sisalana* plants in the Hartwell lab was limited to propagating from the side shoots of a few mature plants only, there was a need to obtain a sufficient quantity of young *A. sisalana* plants in order to permit the design of a fully replicated time course experiment with biological triplicates for each time point, that would form the main experiment of this thesis (Illumina RNA-seq). Fortunately, it was possible to obtain 30 clonal young *A. sisalana* bulbils from the proprietor of the commercial Agave Nursery (Jon Dudek), supplied directly from his nursery in Portugal.

When they first arrived, the thirty *A. sisalana* bulbils measured approximately 20 - 30 cm in length (from base, the lowest part attached to roots, to tip). The young bulbil-staged plants were grown in a greenhouse under a 16 h light/ 8 h dark cycles for 9 weeks. Twenty one plants were selected based on their uniform size and were subsequently entrained in the under controlled 12:12 light/ dark cycles using the Snijders growth chamber (Microclima MC-1000), temperature of 25^{o}C in light and 15^{o}C in dark, relative humidity of 60% in light and 70% in dark, and light intensity of 450 µmole m^{-2} s^{-1} at plant leaf level. This was done 2 weeks prior to sampling in order to allow the plants to entrain to the 12:12 light/ dark cycles. The plants were 11-weeks-old when they were sampled (see Section 2.1.2). The youngest fully expanded leaves of 21 entrained plants were sampled into the following sections: 2-cm-long section from the white basal and pale green basal part (numbered 1 and 2 in Figure 2.2), and 5-cm-long section from tip of leaf (3 in Figure 2.2). In addition, the smallest leaf in the centre of the plant (having peeled away the outer leaves of the central meristematic cone) was also cut in half into an approximately 5-cm-long lower (base; numbered 4 in Figure 2.2) and upper (tip; numbered 5 in Figure 2.2) section (Figure 2.2).

The samples used in this experiment were the same samples that were subsequently selected and used for the main Illumina RNA-seq experiment which was described in subsequent chapters. The samples used included white basal, pale green basal, and dark green tip part of youngest fully expanded leaf, and the tip and base of the smallest leaf in the centre of the plant. For preliminarily determination of the transcript level of a selection of known positive control genes, a single biological replicate was selected from each time point of sampling. The seven plants (numbered 18, 11, 12, 16, 4, 9, and 10) were chosen for each of 7 time points that were sampled at 2 h after lights-on as the first time point and sampled again at 4 h intervals for 24 h with the last sample was collected at the first time point of the next day (2 h after lights-on) respectively (see Table 2.1). The methods for measuring the transcript abundance of each gene of interest, from total RNA extraction to semi-quantitative RT-PCR analysis was primarily the same as the methods used in the initial scoping experiments (Section 3.2.1), except that in each PCR round, only 2 technical replicates were carried out (ideally 3). This was due to the limited number of tubes in PCR well-plate (96) that could not run 105 (7 plants × 5 × samples × 3 technical replicates) samples at the same time. A small selection of the most informative positive control from the preceding scoping experiments, genes representing CAM, the central clock and sucrose metabolism were selected for preliminary analysis in this experiment. These genes included *AsPPC, AsPPDK, GI* and *As_cwINV*and. These were selected as being good examples of CAM genes (*AsPPC, AsPPDK*), a clock gene that peaks at dusk (*AsGI*) and a sucrose metabolism gene (*As_cwINV*) that had been found to be strongly differentially regulated between leaf base and leaf tip, with a large peak of transcript phased to dawn (Figure 3.5E).

Unlike the results from the initial scoping experiments described previously in this chapter (Figs. 3.1 – 3.5), the RT-PCR results obtained from this experiment indicated that the transcript abundance of both *AsPPC* and *AsPPDK* did not differ greatly between the leaf tip and the leaf

leaf base (Figure 3.6A and B). However, the results did distinguish clearly between the leaf tip and the white basal part of the leaf (Figure 3.6A and B). *AsPPC* transcript levels peaked in the middle of the light period in the leaf tip of the youngest fully expanded leaf, and both the base and tip of the youngest leaf sampled from the centre of the apical cone of unexpanded leaves (Figure 3.6A). *AsPPC* transcript levels were generally higher in the leaf tip than the leaf base except at the 10:00 light and 18:00 dark time points, although the transcript abundance of *AsPPC* in the leaf tip was only strikingly higher at 06:00 in the middle of the 12 h light period (Figure 3.6A). The *AsPPC* transcript levels in the white basal part of the youngest fully expanded leaf, and in the base (FH (base); full half of lower or basal part) and tip (FH (tip); full half of upper part or tip) of smallest leaf in the centre of the apical cone were always lower than in leaf tip of the youngest fully expanded leaf (Figure 3.6A). *AsPPC* transcript levels in the white base, FH(base) and FH(tip) were also lower than the pale green basal tissue of the youngest full expanded leaf except that at 06:00, in the middle of the light period, the level of *AsPPC* transcripts in the FH(tip) was slightly higher than the level in the pale green base of the youngest fully expanded leaf (Figure 3.6A).

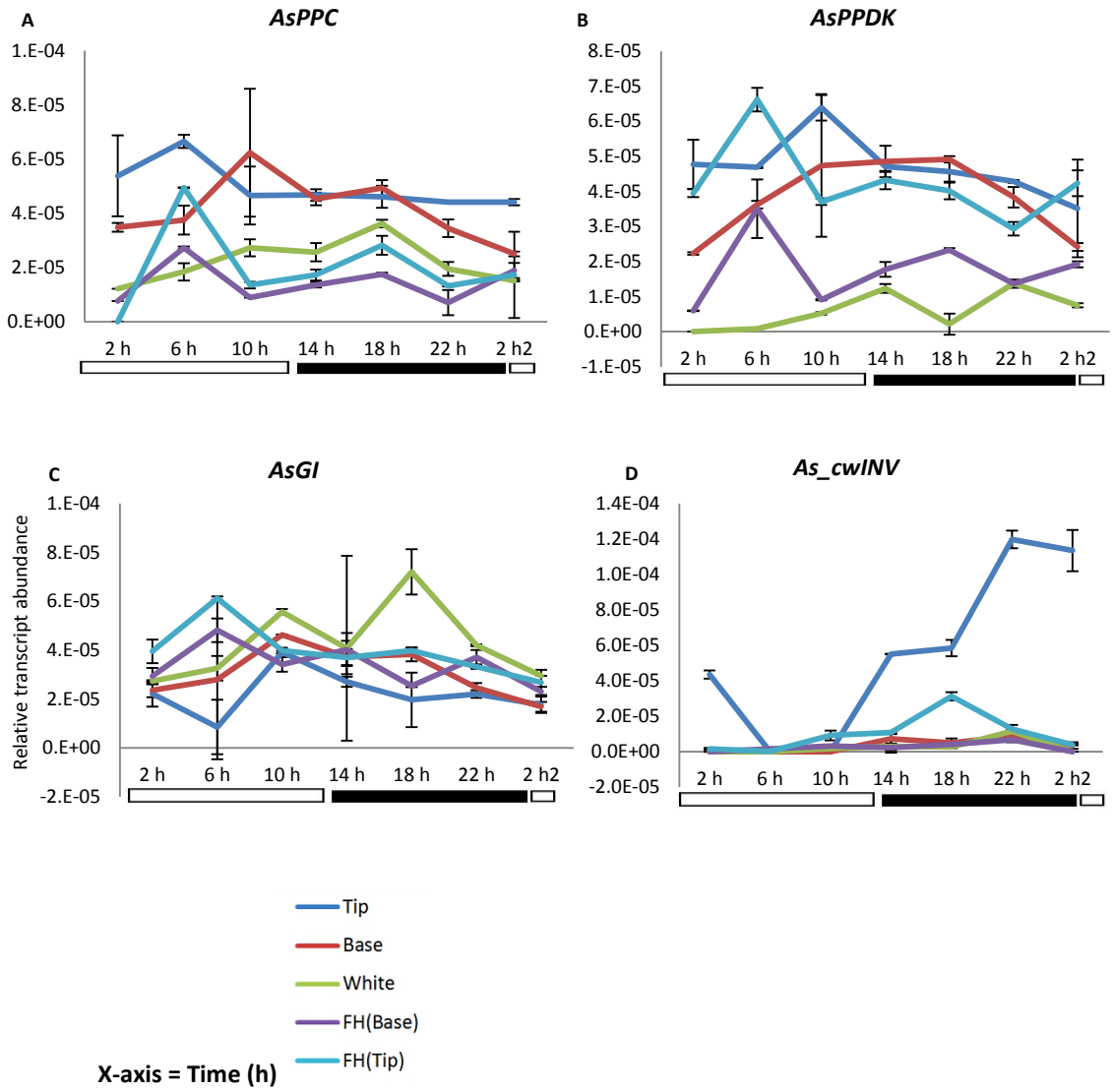


Figure 3.6 Relative transcript abundance level of CAM, circadian clock gene and fructan related gene.

CAM genes *AsPPC* (A) and *AsPPDK* (B), circadian clock gene *AsGI* (C) and sucrose related gene *As_cwINV* (D) in dark green tip, pale green base, and white basal part of youngest fully expanded leaf and tip (FH(Tip)) and base (FH(Base)) of smallest leaf in the centre of 11-week-old plants sampled at 4 h intervals throughout 12:12 light/ dark cycles. 2 h2 was the last sample collected and was a replicate of the first time point, but was collected on new, final day of the experiment (2 h after lights-on). As such it is a repeat of the 2 h sample, but was collected 24 h later. Only 2 technical replicates were used (ideally 3). This was due to the limited number of tubes in PCR well-plate (96) that could not run 105 (7 plants × 5 × samples × 3 technical replicates) samples at the same time. Transcript abundance was determined using semi-quantitative RT-PCR; values were normalized to abundance of UBQ10 transcripts.

AsPPDK transcript levels were variable between all of the leaf sections and time points collected in this experiment. The transcript level of *AsPPDK* was mostly higher in the leaf tip than the leaf base of the youngest fully expanded leaf, except at the 14:00 dark and 18:00 dark time points (Figure 3.6C). In the FH(Tip) of the youngest unexpanded leaf within the apical cone of leaves, the transcript level of *AsPPDK* was similar to the transcript level in the leaf tip and base of the youngest fully expanded leaf (Figure 3.6B). The very young FH(tip) sample actually reached a slightly higher peak than the tip sample from the youngest fully expanded leaf at 06:00 light (Figure 3.6B).

The *AsGI* and *As_cwINV* genes were analysed as positive control genes for circadian clock control and light/dark regulation. *AsGI* transcript levels in the tip and base of the youngest fully expanded leaf peaked at 10:00 light (Figure 3.6C). FH(base) and FH(tip) both peaked for *AsGI* at 06:00 in the middle of the 12 h light period, but the white basal tissue from the youngest fully expanded leaf displayed a *GI* transcript peak at 18:00 dark, with a second minor peak at 10:00 light (Figure 3.6C).

As_cwINV provided the most striking and clear result out of all of the genes analysed (Figure 3.6D). Its transcript abundance was either very low or undetectable at 06:00 light and 10:00 light, increased dramatically throughout the dark period, and peaked at the end of the dark period at 22:00 dark (2 h before dawn) in leaf tip of youngest fully expanded leaf (Figure 3.6D). *As_cwINV* transcripts were also detected in the tip of the youngest unexpanded leaf in the centre of the apical cone, namely in the FH(Tip) samples, reaching their 24 h peak in the middle of the dark period (Figure 3.6D). By contrast, *As_cwINV* transcript levels were either very low or undetectable in all of the other leaf samples and time points (Figure 3.6D). Thus, *As_cwINV* was found to be a strong marker gene for the leaf tip of the youngest fully expanded leaf, but also showed a development gradient of its transcript abundance along the length of

the very youngest unexpanded leaf within the central apical leaf cone. It should be noted that *As_cwINV* transcript level in leaf tip at 2h2 (2 h after lights-on of the following day) was considerably higher than the transcript level of the leaf tip sampled at 2 h (2 h after lights-on of the first sampling day) while in other genes the transcript levels from these 2 time points were approximately the same, which was made sense (Figure 3.6D). However, the size of the error bar for this time point (2h2) was the largest compared to the other time points, which might make the interpretation of the result for this time point (2h2) more error prone, relative to other time points. While the result for this time point was not ideal, and should be repeated further, this result for *As_cwINV* transcript levels still provided valuable information as the other time points showed a very consistent result compared to the previous results for *As_cwINV* using older, 1.5-year-old *A. sisalana* plants (Figure 3.5E).

3.3 Summary

One of the main goals of this PhD project was to perform a transcriptome-wide analysis of the genes involved in CAM in *A. sisalana* using high-throughput sequencing (Illumina RNA-seq). It was not feasible due to funding limitations to sequence all of the RNA samples from all tissues and time points. Instead, the approach used here was to identify the optimal leaf-sampling regime for the RNA-seq work, the optimal leaf samples being those that were most likely to facilitate the identification of the CAM-associated genes. The proposed plan was to obtain the assembly of leaf transcriptome sequenced from the CAM leaf tissues (leaf tip) and C_3 tissues (leaf base). These different tissues were expected to display significant differences in terms of transcript abundance level of CAM genes. To test these ideas, the initial scoping study described in this chapter was an essential task in order to guide the appropriate selection of the most suitable leaf segments and time points for the subsequent RNA-seq work.

The series of preliminarily transcript abundance scoping experiments described in this chapter employed previously well-known CAM, central clock and fructan/ sucrose metabolism associated genes as marker genes to test both for transcript abundance differences between CAM and non-CAM leaf tissues, and for the regulation of the transcript levels over the light/ dark cycle. In general, there was considerable variability for a number of these initial semi-quantitative RT-PCR scoping experiments. Whilst this might be a tissue specific phenomenon, the careful selection of leaf tissue material for further quantitative analysis of the regulation of transcripts required critical consideration. In the initial scoping experiments, *AsPPC* and *AsPPDK* were mostly consistent with the previous studies in terms of their transcript abundance; increasing markedly in the leaf tip where CAM was proposed to be localised, compared to the leaf base where C_3 was proposed to predominate. However, results from the scoping experiment using the same samples used in for the later detailed RNA-seq experiment (chapter 5) showed that both *AsPPC* and *AsPPDK* transcript abundances were not significantly distinguished between leaf tip and leaf base. However, they were clearly distinguished between leaf tip and the white basal part of the youngest fully expanded leaf from the 11-week-old plants (Figure 3.6A and B). Hence, the white basal part of the leaf was selected to be included in the main Illumina RNA-seq experiment as it was the best 'non-CAM' leaf tissue available from all of the initial scoping experiments (Figure 3.6).

Based on the findings for *AsPPC* and *AsPPDK* in the initial scoping experiments which showed a clear difference in transcript level between leaf tip and leaf base (Figure 3.5A and C) and the results of the preliminary experiments using the total RNA samples that were subsequently used in the main Illumina RNA-seq samples showing a clear difference between leaf tip and white basal part described above (Figure 3.6A and B), these leaf tip, base and white basal part of the leaf samples were considered to be suitable for taking forward into the Illumina RNA-seq experiment. Specifically, the samples selected for the RNA-seq experiment were the white

basal, pale green basal, and dark green tip sections of the youngest fully expanded leaf sampled at both 10:00 light (2 h before dusk) and 22:00 dark (2 h before dawn). The selection of the time points was based on the finding that transcript level of *As_cwINV* was lowest at 10 h and peaked at 22 h, and *AsGI* transcript peaked at 10 h in the majority of tissues sampled. Due to the high-cost of the RNA-seq analysis combined with a limited available budget, a pragmatic decision was taken to sequence just these two representative light and dark samples. The remaining samples spanning the rest of the 24 h cycle (02:00 light, 06:00 light, 14:00 dark and 18:00 dark) were stored safely in the -80˚C freezer so that they could be used for later detailed Q-RT-PCR analysis of genes identified as potential CAM-associated genes through the RNA-seq analysis.

Chapter 4

An investigation of the developmental and light/ dark regulation of CAM along the length of young *A. sisalana* leaves: physiological and biochemical analysis of CAM-associated characteristics

4.1 Introduction

The extent to which CAM is occurring and being utilised within a plant leaf or stem/ cladode can be defined and characterised using a number of key physiological and biochemical measurements for traits that associate strongly with CAM. These include nocturnal CO_2 fixation by the leaf, the associated daily cycling of the major pools of leaf malate and storage carbohydrates, and the induction and light/ dark regulation of key CAM-associated enzymes, transporters and their cognate regulatory proteins such as *PPCK* (Borland *et al.*, 2009; Dever *et al.*, 2015; Yang *et al.*, 2015). A wide-array of previously published studies has investigated CAM species with respect to these and other key parameters associated with CAM. For example, Chen *et al.*, (2002) performed a comparative study of the diurnal changes in metabolite levels in the leaves of 3 different CAM species, namely *Ananas comosus* (pineapple)*, Kalanchoë daigremontiana* and *K. pinnata.* They reported that there were changes in the levels of several key metabolites associated with central metabolism, especially the levels of the main organic acids, hexose-phosphates, and oxaloacetate, PEP and pyruvate, all of which correlated with CAM in these species (Chen *et al.*, 2002). All 3 CAM species exhibited the classic CAM feature

of a nocturnal accumulation of malate, reaching a daily peak at dawn, but in addition, several of the other measured metabolites also cycled in abundance over the light/ dark cycle.

Nocturnal CO_2 fixation has for many decades been the most intensively studied physiological correlate for CAM; with the daily cycle of CAM often being described and deciphered with respect to the classic 4–phase framework for the CAM pathway (Osmond, 1978). In addition to the light/ dark cycle of CAM CO_2 fixation, studies using constant environmental conditions (constant light, temperature and humidity) have demonstrated that there is an endogenous circadian rhythm of CO_2 fixation in CAM leaves of *K. fedtschenkoi* and *K. daigremontiana,* which persists with peaks and troughs of CO_2 fixation for many days in constant light and temperature conditions (Wilkins, 1992; Hartwell, 2006). Based on these robust circadian rhythms of CO_2 exchange in *Kalanchoë* species, it is believed that the circadian clock plays a fundamental role in forming and synchronizing the temporally divided metabolic components of CAM (Wilkins, 1992; Hartwell, 2006; Lüttge, 2000).

A study of the daily carbon cycle in leaves of the monocot CAM crop species *A. comosus* suggested that soluble sugars play an important role in the provision of carbohydrates for glycolytic provision of PEP for PEPC in the dark period during CAM (Kenyon *et al.*, 1985). Sucrose was proposed to be the main soluble storage sugar in CAM leaves followed by fructose and glucose (Kenyon *et al.*, 1985; Wang and Nobel, 1998). However, Antony *et al.*, (2008) demonstrated that nocturnal turnover of starch in the leaves of two varieties of pineapple was sufficient to provide all of the PEP required for the dark period CO_2 fixation. These authors also demonstrated that soluble sugars were turned over during the dark period in both pineapple cultivars (Antony *et al.*, 2008). Thus, precise determination of the exact source of carbohydrate used for PEP provision in pineapple must await further more detailed research, perhaps using molecular genetic approaches to block either starch or soluble sugar

turnover in order to dissect which pool of leaf carbohydrates is most important for PEP provision for CAM.

In Agaves, which have evolved CAM completely independently from *Kalanchoë* and pineapple, a further potential source of leaf carbohydrates for PEP provision in the dark are the leaf fructans that are found in many species of Agave as the major storage carbohydrate in stems, along with other carbohydrates associated with its metabolism (Mancilla-Margalli and Lopez, 2006). It has been proposed that sucrose is the storage carbohydrate for glycolysis to supply the PEP for nocturnal atmospheric CO_2 fixation by PEPC in Agave species (Christopher and Holtum, 1996; Wang and Nobel, 1998). Wang and Nobel, (1998) demonstrated that sucrose, glucose and fructose accounted for 98 % of the soluble sugars found in the photosynthetic chlorenchyma cells of *A. deserti*, whereas the only fructan detected in the chlorenchyma (where CAM occurs) was neokestose (3- degrees of polymerisation), which only accounted for 2 % of soluble sugars in the chlorenchyma. They further reported that sucrose in the chlorenchyma was sufficient to produce malate through dark CO_2 fixation as one sucrose could be used to produce four malates through being broken down through glycolysis to 4 PEP which were then utilised in CO_2 fixation by PEPC. Thus, they proposed that *A. deserti* used sucrose as its source of carbohydrates for CAM CO_2 fixation in the dark (Wang and Nobel, 1998).

In CAM species such as *K. fedtschenkoi* and *M. crystallinum*, a range of experimental approaches have revealed that the transcript abundance, level of translatable mRNA, and activity of PPCK, and the resulting phosphorylation state and apparent K_i for malate of PEPC increase in the dark period coincident with increased flux through PEPC for dark CO_2 fixation (Hartwell *et al.*, 1999; Taybi *et al.*, 2000; Taybi *et al.*, 2004). Little if any equivalent data for the regulation of *PPCK* RNA levels and activity, and PEPC phosphorylation state or even the apparent K_i of PEPC for malate are available for monocot CAM species such as pineapple,

Agave or Aloe. (Theng *et al.*, 2008) did provide preliminary evidence for dark phosphorylation of PEPC in leaves of pineapple by using a ProQ Diamond phosphoprotein gel stain on SDS-PAGE gels to identify a band of the same molecular weight as PEPC that increased in phosphorylation level in the dark period. They also reported that the % inhibition of PEPC by L-malate in pineapple leaves decreased in the dark period suggesting that PEPC in pineapple was becoming phosphorylated and thus less sensitive to malate inhibition in the dark, coincident with dark CO_2 fixation associated with CAM (Theng *et al.*, 2008). Furthermore, a study of CAM in the epiphytic, atmospheric Bromeliad *Tillandsia pohliana*, which represents a further independent origin of CAM in the monocots, revealed that PEPC became less sensitive to malate inhibition in the dark, with a distinct light/ dark pattern of regulation for the K_i of PEPC for feedback inhibition by malate (Freschi *et al.*, 2010). However, no evidence relating to light/ dark regulation of PEPC is currently available for any Agave species. Indeed, despite an increasing number of related studies being published in recent years, many other Agave CAM-associated metabolic and physiological features relating to the light/dark regulation of CAM, and CAM development along the length of the leaf remain to be studied in any detail even in a single Agave species (Matiz *et al.*, 2013).

In the results presented in this chapter, the same *A. sisalana* leaf section samples used for the RNA-seq analysis described in chapters 5, 6 and 7 were also used for sugar and malate level determination, and immuno-blot analysis of the abundance of key CAM proteins, including the use of an anti-phospho-PEPC antibody to investigate the phosphorylation status of PEPC over the light/ dark cycle in *A. sisalana* leaves. CO_2 gas exchange experiments performed previously by Dr. Susanna Boxall and Dr. James Hartwell at the University of Liverpool are also presented and discussed in this chapter in order to provide the leaf physiological context for the metabolite and protein abundance measurements. In addition, *A. sisalana* leaf tip samples collected every 4 h for 82 h from plants placed under free-running conditions of constant light,

temperature and humidity conditions (LL; light 100 µmoles m^{-2} s^{-1}, temperature 15 ˚C, 60 % humidity) were analysed in order to investigate whether leaf malate and sugar levels were under circadian clock control. It had also been the intention to analyse the activity of key CAM enzymes as part of the study of the same samples used for the later RNA-seq analysis. However, a combination of time constraints and limited remaining sample sizes prevented this aspect of the study from being completed. Despite this, the presented immuno-blot results for several CAM proteins provided a complementary dataset, which revealed the amount of each CAM enzyme even though it was not possible to determine their activity within this study. Together with the preliminary scoping experiments that investigated the regulation of the transcript abundance of CAM, circadian clock and sugar metabolism associated genes along developing leaves of *A. sisalana* (Chapter 3), the experiments presented in this chapter, which provided data for various metabolites and physiological characteristics associated with CAM, were aimed to build supporting information concerning the level of CAM along the leaf development longitudinal axis, and the light/dark regulation of CAM in the different leaf segments. The combined results from these experiments thus provide a robust framework of correlated data to underpin the analysis and interpretation of the RNA-seq experiment described in chapters 5, 6 and 7.

4.2 Result and discussion

4.2.1 Gas exchange analysis of CAM and its light/ dark and circadian clock control in developing leaves of A. sisalana

Unfortunately, it was not possible to undertake leaf gas exchange experiments on different sections of the *A. sisalana* leaves during this PhD project due the fact that the multi-cuvette infra-red gas analyser (IRGA) system was broken for much of the last two years of this PhD.

However, luckily, preliminary leaf CO_2 exchange results for *A. sisalana* leaves had been obtained previously by Dr. Susie Boxall and Dr. James Hartwell, University of Liverpool, and those results are presented here in order to exemplify the patterns of gas exchange associated with different segments of the developing *A. sisalana* leaf, and also to highlight the regulation of CAM-associated CO_2 fixation under constant light, temperature and humidity conditions. The first experiment was conducted using different segments of the youngest fully expanded *A. sisalana* leaf. The leaf was cut into a number of segments along its length as shown in the photograph to the right of Figure 4.1, and the basal end of each leaf segment was submerged in distilled water in the small beakers that were subsequently placed in the gas exchange cuvettes of the multi-cuvette IRGA system. The distilled water beakers were covered with several layers of parafilm in order to prevent evaporation or respiratory gases from the distilled water from influencing the gas exchange data. CO_2 fixation was measured approximately every 10 mins for each leaf section over several cycles under 16:8 light/ dark conditions. A representative 24 h period of the gas exchange data is presented (Figure 4.1). The gas exchange values were corrected for the leaf area in each cuvette by measuring the leaf area of each leaf segment at the end of the experiment and using a recalculation feature of the gas exchange software to recalculate the level of photosynthetic CO_2 fixation on a per m^2 of leaf tissue basis. The results demonstrated very clearly that the leaf base section was performing predominately C_3 photosynthesis as it fixed atmospheric CO_2 mainly during the light period (Figure 4.1). However, the leaf base did achieve very low levels of net atmospheric CO_2 fixation in the dark period, suggesting perhaps that the oldest chlorenchyma cells at the tip-end of the basal leaf section may have been capable of some dark CO_2 fixation. This emphasises that these were relatively large sections of the leaf (5 – 6 cm in length) which would of course have included many thousands of cells of different developmental ages, and

that the CO_2 exchange pattern is an average for all of the photosynthetically competent chlorenchyma cells in each leaf section (Figure 4.1).

In strong contrast to the gas exchange data for the leaf base, the leaf tip was found to be performing full CAM, with all four phases of CAM clearly evident and a peak dark CO_2 fixation rate of almost 5 µmoles m^{-2} s^{-1} (Figure 4.1). However, even the leaf tip section that performed the most dark CO_2 fixation only reached 0 µmoles m^{-2} s^{-1} of net atmospheric CO_2 fixation in the light (full phase III; refixation of CO_2 generated through malate decarboxylation behind closed stomata) for a single 10 minute time point, approximately 2 h after lights on (Figure 4.1). However, the shape of the gas exchange curve for the leaf tip section in the light did match well with phases II, III and IV of CAM, with the phase III period of low atmospheric CO_2 fixation in the light lasting for approximately 4 h from around 08:00 until around 12:00 (6 h into the 16 h light period; Figure 4.1). The leaf tip displayed a large phase IV throughout the remaining 10 h of the light period (12:00 until 22:00) revealing that even the mature leaf tip was able to perform direct C_3-type fixation of atmospheric CO_2 for the majority of the light period. This may be a consequence of a number of factors, including the fact that the plants were well watered prior to leaf excision for this experiment, plus the cut basal end of each leaf section was submerged in distilled water, such that the leaf section most probably had an unlimited water supply. Thus, the gas exchange pattern is representative of the photosynthetic physiology of each leaf section under fully-hydrated, well-watered conditions. Considering that *A. deserti* has been reported to convert to performing full C_3 photosynthesis under well-watered conditions (Hartsock and Nobel, 1976), it is perhaps not surprising that these young leaf sections of well-watered *A. sisalana* leaves were able to perform a large amount of phase IV C_3-type atmospheric CO_2 fixation through the direct action of Rubisco in the Calvin cycle, even for the mature leaf tip (Figure 4.1). The findings for the gas exchange pattern of the leaf tip of these young *A. sisalana* plants are entirely consistent with the gas exchange pattern

reported for the closely-related species *A. angustifolia*, which also displayed a large phase IV period of atmospheric CO_2 fixation in the light period (Holtum and Winter, 2014; Winter *et al.*, 2014).

As the leaf sections increased in age from the most basal section to the tip, there was an obvious transition from C_3 photosynthesis at the leaf base to CAM at the leaf tip (Figure 4.1). Nocturnal CO_2 fixation increased, whereas fixation during the light decreased as the leaf sections matured from base to tip (Figure 4.1). Interestingly, the pattern of CO_2 fixation in the dark displayed a clear dip in dark CO_2 fixation in the middle of the 8 h dark period for the two middle sections of the leaf (green and purple lines, Figure 4.1). It is worthwhile to speculate that this dip may relate to a delay in the activation of PPCK and the resulting phosphorylation of PEPC in these middle sections of the leaf, and it would be fascinating to test this idea further in the future using immunoblots to study the phosphorylation state of PEPC throughout the dark period in these middle leaf sections.

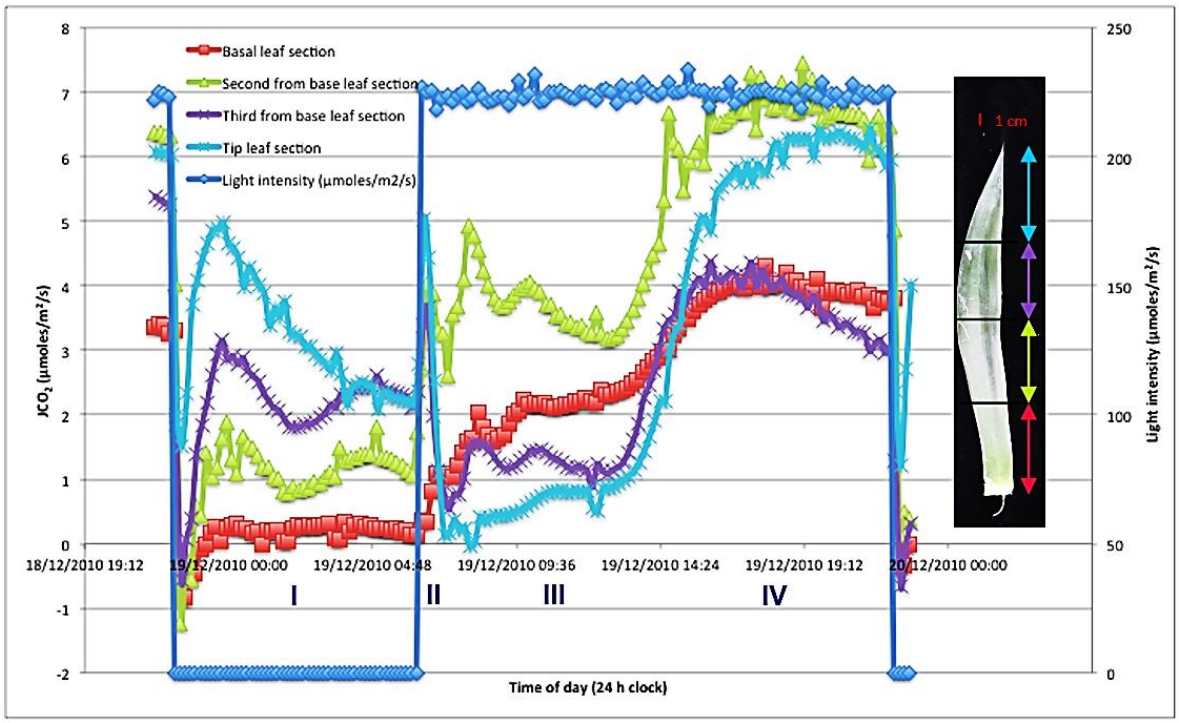


Figure 4.1 The 24 h light/ dark gas exchange pattern of different leaf sections from a young, fully expanded *A. sisalana* leaf reveal a developmental progression along the leaf from C_3 at the base to full CAM at the tip.

The graph show the 24 h net CO_2 exchange pattern for different sections of the youngest fully expanded leaf from a young *A. sisalana* plant (mature leaves 20 – 25 cm long). Basal section (red), second from base (green), third from base (purple), and tip section (blue) of a leaf excised from an *A. sisalana* plant were used to determine CO_2 fixation under 16:8 light/ dark conditions. Light intensity (blue diamonds) is plotted against the right-hand y-axis. Roman numbers indicate the phase of CAM.

In order to investigate whether or not CAM-associated nocturnal CO_2 fixation is under the control of the circadian clock in *A. sisalana*, the leaf tip section that had already been found to have the highest level of dark CO_2 fixation, and thus CAM, was used for a gas exchange experiment in which gas exchange was measured over several 12:12 light/ dark entrainment cycles prior to transfer to constant light, temperature and humidity (LL) free-running conditions. LL conditions test for the occurrence of the free-running circadian rhythm of CO_2 exchange associated with CAM, which has been reported previously and studied in detail in the dicot, obligate CAM species, *K. fedtschenkoi* and *K. daigremontiana* (Wilkins, 1992; Hartwell, 2006; Lüttge, 2000). A CAM CO_2 fixation rhythm has also been reported for the

inducible CAM species *M. crystallinum* when it is performing CAM (Dodd *et al.*, 2003). A free-running circadian rhythm of CO_2 exchange had not been reported previously for any monocot CAM species, and so this experiment was important in order to assess whether or not control via the central circadian clock was likely to be an important aspect of the daily regulation of CAM in *A. sisalana*.

In the *A. sisalana* leaf tip, the strong CAM pattern of light/ dark CO_2 fixation was again observed in this second gas exchange experiment, with dark CO_2 fixation reaching a peak at over 8 µmoles m^{-2} s^{-1}, and a long and substantial period of phase IV CO_2 fixation also present (Figure 4.2). Phase III was more pronounced and prolonged in this leaf tip from a mature leaf, and the phase II burst of CO_2 fixation at the beginning of the light period was very short lived (Figure 4.2).

When this leaf tip section from a mature leaf was subsequently subjected to constant light (LL) conditions for 4 days, it was observed to exhibit a robust, free-running circadian rhythm of CO_2 fixation that persisted throughout the LL conditions (Figure 4.2). The rhythm was characterised by periodic peaks and troughs in the CO_2 fixation pattern, which initially coincided with subject dark (peaks) and subjective light (troughs), but very rapidly phase advanced ahead of their expected timing due to the short period of the CO_2 fixation rhythm under LL conditions (Figure 4.2). This *A. sisalana* CAM CO_2 fixation rhythm correlated well with the previously studied short period CAM CO_2 fixation rhythms reported for CAM leaves of *K. fedtschenkoi* and *K. daigremontiana* (Hartwell *et al.*, 1996; Hartwell *et al.*, 1999; Nimmo, 2000).

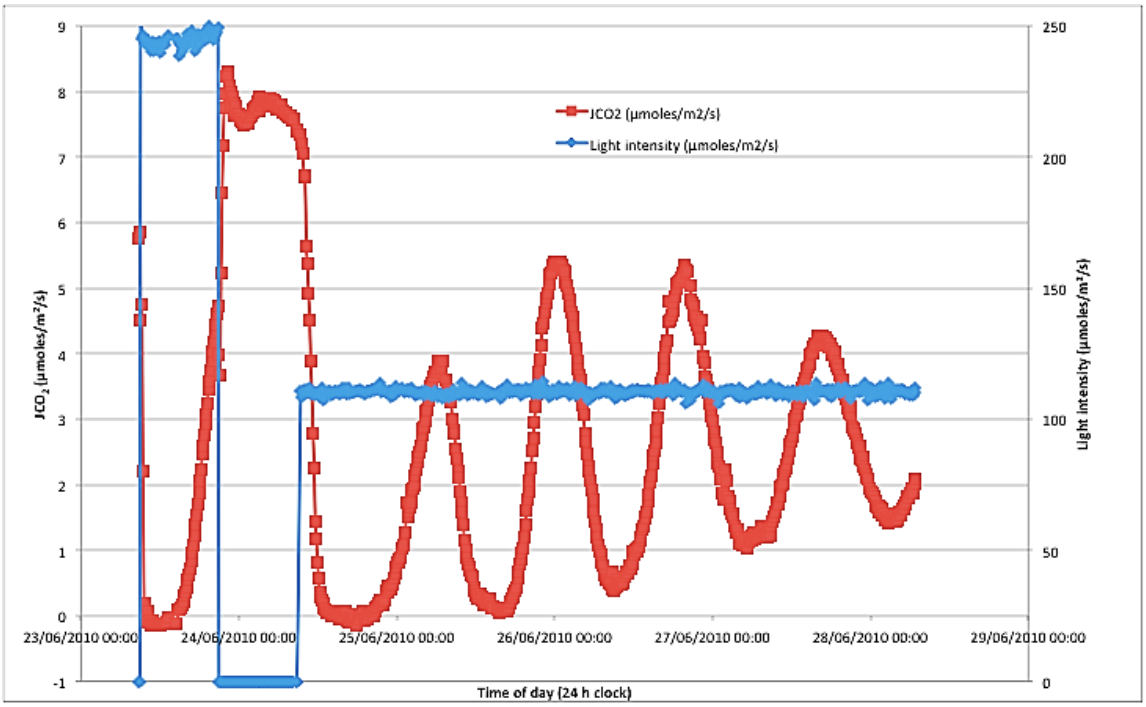


Figure 4.2 The leaf tip section of a mature *A. sisalana* leaf displays a robust circadian rhythm of CO_2 exchange under constant light, temperature and humidity conditions.

CO_2 fixation was measured from the mature *A. sisalana* leaf tip under 12:12 light/dark (LD) entrainment conditions prior to release into free-running constant light (LL) conditions for 4 days. Left-hand y-axis indicates CO_2 fixation. Light intensity (blue diamonds) is plotted against the right-hand y-axis.

4.2.2 Protein Abundance Determination

Several of the enzymes and membrane localised metabolite transporters required for the efficient operation of the CAM pathway are known, even for Agaves. For example, CAM plants use PEPC to fix CO_2 in the dark, and in CAM species such as *K. fedtschenkoi* or *M. crystallinum*, this is known to be activated in the dark by phosphorylation by the protein kinase PPCK. At dawn, accumulated malic acid is released from the vacuole into the cytosol. It is then decarboxylated by either NAD- and/or NADP-malic enzyme (ME) or PEP carboxykinase (PEPCK), depending on the plant species (Dittrich, 1976). In the case of ME-type CAM species, pyruvate and CO_2 are produced as the products of malate decarboxylation in the light. CO_2 is refixed in the light by Rubisco in the C_3 Calvin cycle, and pyruvate is transformed to PEP by

PPDK (Hartwell, 2006). The abundance of these key CAM proteins can be a valuable and informative marker for the identification of tissues that perform CAM. To assess the abundance of the CAM proteins PEPC, PPDK and the α and β-subunit of NAD-ME, plus the phosphorylation state of PEPC, a number of immune-blots were performed using specific antisera for each protein or phopho-protein.

In order to achieve even loading of the same amount of total protein in each well of the SDS-PAGE gel, the concentration of total protein in each leaf extract was determined using a Bradford assay (see chapter 2 for method details). The total protein concentration for the extracts prepared for the leaf tip samples was consistently higher than for the leaf base sections, even though the fresh weight of the samples was the same prior to extraction. This may be a consequence of the more succulent tissue at the leaf base where a greater proportion of the cells are likely to be the large water storage parenchyma in the centre of the leaf. Thus, a larger proportion of the fresh weight at the leaf base was likely to be water. However, these differences in total protein should not have affected the immune-blot results as each well of the SDS-PAGE gels was loaded with the same amount of total protein (see chapter 2 for details).

Staining of replica SDS-PAGE gels with Coomassie Blue confirmed that the total protein loaded into each well was even from well to well (data not shown). All of the antibodies used in this study had been obtained previously from other labs based on published reports, and they had all been used successfully to detect the corresponding CAM proteins in leaves of *K. fedtschenkoi* (e.g. Dever *et al*., 2015). However, antibodies to NADP-ME and the phosphorylated form of PPDK did not work for *A. sisalana* (data not shown).

The protein abundance for each of the probed proteins was determined using samples from the dark green tip, pale green base and basal white segment of the youngest, fully-expanded

A. sisalana leaf sampled at 10:00 (light, 2 h before dusk) and 22:00 (dark, 2 h before dawn). The samples used were from the young *A. sisalana* plants used for the RNA-seq experiment and had been grown and sampled under 12:12 light/dark conditions using the Snijders Microclima growth cabinet. Only the leaf tip, pale green basal and white basal tissue sampled at 10:00 and 22:00 were used for the immune-blotting experiments due to the limited number of wells on the SDS-PAGE gels, plus these were the same samples and time points used in the RNA-seq experiment descried in detail in Chapter 5.

The results showed clearly that each CAM protein measured was most abundant in the leaf tip samples, with PEPC and β-NAD-ME also detected with a weak intensity band in the pale green base and white basal section of the leaf (Figure 4.3). This correlated well with the data presented at the end of chapter 3 using the same leaf samples, which demonstrated that *AsPPC* and *AsPPDK* transcript levels were higher in the leaf tip than the leaf basal white tissue throughout the light/ dark cycle (Figure 3.6A and B). However, the transcript data showed only relatively small differences in the levels of the corresponding *AsPPC* and *AsPPDK* transcripts between the pale green basal part of the leaf and the dark green tip (Figure 3.6A and B), whereas the level of the corresponding proteins was extremely low to undetectable in the pale green base, especially relative to the intense band obtained for the dark green leaf tip (Figure 4.3). The protein abundance results for α- and β-NAD-ME also support the strong increase in the abundance of CAM-associated proteins in the leaf tip relative to the pale green base and white basal tissue of the leaf (Figure 4.3). Overall, these immuno-blot results further emphasise strengthen support for CAM occurring predominantly in the mature leaf tip of the youngest, fully-expanded leaf of young *A. sisalana* plants. These key CAM proteins have also been detected as correlating strongly and specifically with strong CAM in mature leaves of the model CAM species, *K. fedtschenkoi* (Dever *et al.*, 2015).

The results for the immuno-blot analysis of the level of the phosphorylated form of PEPC provided a particularly clear and informative result (Figure 4.3). Phospho-PEPC was only detected in the leaf tip in the dark sample at 22:00, all other lanes of the immuno-blot were blank (Figure 4.3). This revealed for the first time that PEPC in the CAM performing segment of *A. sisalana* leaf is subject to regulatory phosphorylation in the dark. Based on published work in other CAM species (e.g. Hartwell *et al*., 1999; Dever *et al*., 2015), it is highly likely that this dark period phosphorylation of PEPC in the *A. sisalana* leaf during CAM resulted in an increase in the K_i of the PEPC for feedback inhibition by malate during the dark period, in turn allowing PEPC to remain active in the face of the mounting concentration of malate that accumulates throughout the dark period, and thus allowing dark period CO_2 fixation via PEPC to proceed in an optimal manner for a greater proportion of the dark period, resulting in a greater accumulation of malic acid by dawn.

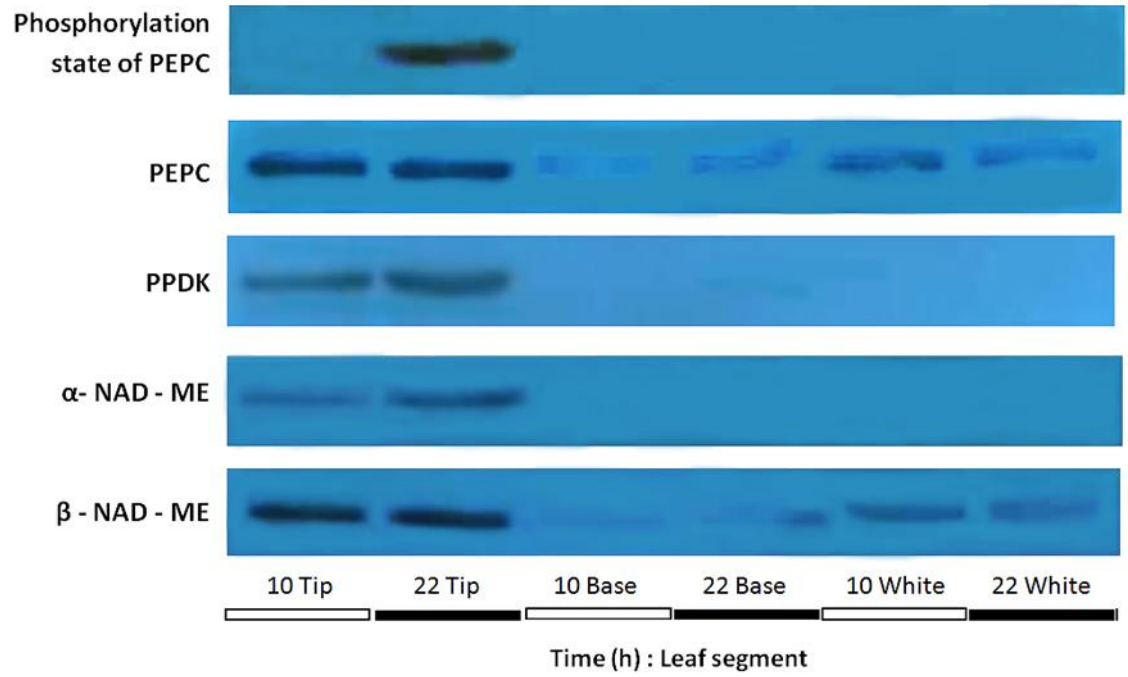


Figure 4.3 Immuno-blot analysis demonstrates that CAM proteins are almost exclusively detected in the leaf tip in *A. sisalana*, and that PEPC is phosphorylated only in the dark in the leaf tip.

Immuno-blots were used to measure the level of phospho-PEPC, PEPC, PPDK, α- and β NAD-ME using *A. sisalana* leaf samples from the youngest, fully-expanded leaf from 11-week-old plants. Samples used included dark green tip, pale green base and white basal section sampled at 10:00 (light, 2 h before dusk) and 22:00 (dark, 2 h before dawn). Plants were entrained under 12:12 light/ dark cycles in a Snijders Microclima growth cabinet for two weeks to achieve full entrainment prior to sampling.

As a full 12:12 time course had been collected for the leaf tip as part of the main experiment used for the RNA-seq described in chapter 5, the leaf tip samples collected at 4 h intervals were used investigate the light/dark regulation of abundance of the CAM proteins (Figure 4.4). The results revealed that phosphorylation state of PEPC increased during the dark period; with all three dark samples, collected at 14:00, 18:00 and 22:00, displaying a high level of PEPC phosphorylation (Figure 4.4). PEPC was not detected as being phosphorylated for any of the light period samples (Figure 4.4). This result correlated perfectly with previously published work which demonstrated that *AsPPCK* transcript abundance, translatable mRNA and activity, and also the phosphorylation state of PEPC all peaked at the middle of the night and were undetectable during the day in *K. fedtschenkoi* (Hartwell *et al.*, 1999; Dever *et al.*, 2015). In

CAM-induced *M. crystallinum* leaf, it was also found that the transcript abundance of *AsPPCK* was higher in the dark than in light time (Taybi *et al.*, 2000; Boxall *et al.*, 2005). This, together with the result in Figure 4.3, indicates that the phosphorylation state of PEPC is likely to be light/ dark regulated in *A. sisalana*, with a temporal pattern of regulation that is very similar to the pattern reported previously in dicot CAM species that have evolved CAM completely independently from *A. sisalana*.

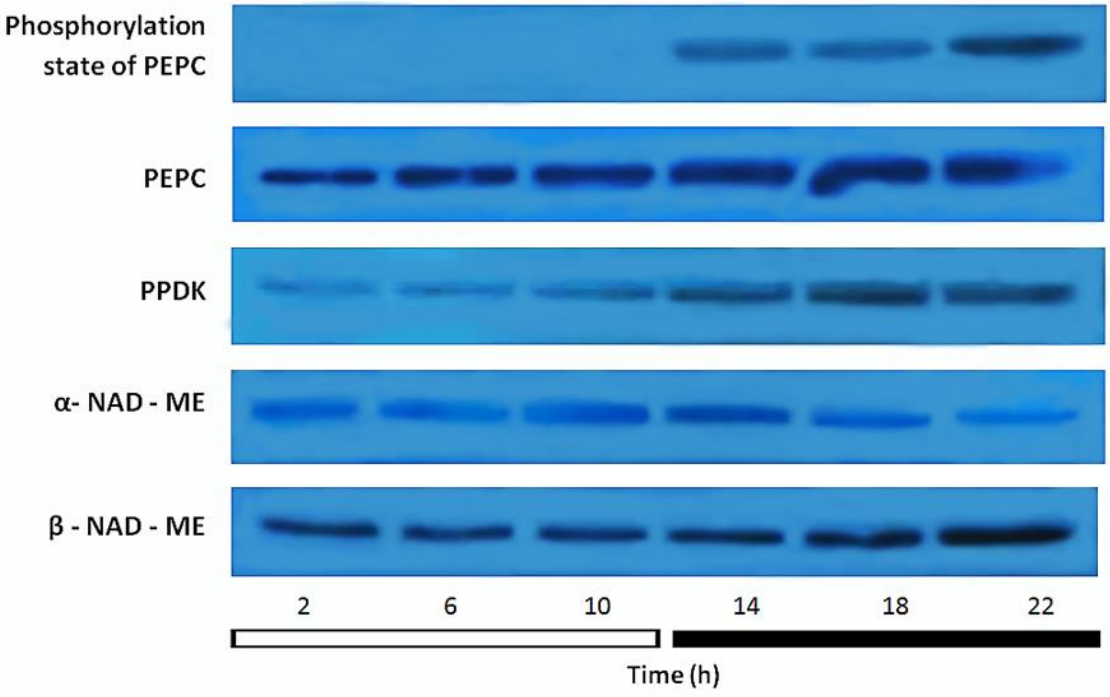


Figure 4.4 Immuno-blot analysis demonstrates light/dark regulation of abundance of the CAM proteins in *A. sisalana* leaf tip.

Immuno-blots were used to measure the level of phospho-PEPC, PEPC, PPDK, α- and β NAD-ME using *A. sisalana* leaf samples from the youngest, fully-expanded leaf from 11-week-old plants. Samples used included green tip sampled at 4 h intervals throughout a 12:12 light/dark cycle where plants were entrained in a Snijders Microclima growth cabinet for two weeks to achieve full entrainment prior to sampling.

The rest of CAM enzymes measured had relatively constant protein abundance levels throughout the light/ dark cycle in the *A. sisalana* leaf tip (Figure 4.4). However, there did appear to be small increases in the level of PPDK and β-NAD-ME in the dark period, although

this may be due to minor variation in the loading of total protein in each well of the gel. The protein abundance of α-NAD-ME and PPDK was noticeably lower than the other CAM proteins. However, the immuno-blot procedure includes a number of technical variables that will lead to variation in the intensity of the bands on the final X-ray film exposure. In particular the duration of the exposure of the blot to the X-ray film can produce marked differences in the intensity of the bands.

The light/ dark changes in the transcript level of PEPC determined in chapter 3 (Figure 3.5A and 3.6A) were not reflected by changes in PEPC protein abundance, which was found to be present at a constant level over the entire light/ dark cycle through the immuno-blot analysis (Figure 4.4). This suggests that the protein abundance of PEPC was stable over the light/ dark cycle and hence that PEPC was not rapidly turned over during the light/ dark cycle. However, it remains possible that the level of active and phosphorylatable PEPC varied over the light/ dark cycle as it is not known where the epitope targeted by the PEPC antibody resides within the protein sequence.

The protein abundance of PPDK, and α- and β-NAD-ME, also varied little over the light/ dark cycle (Figure 4.4), whereas the transcript abundance of PPDK varied over the light/ dark cycle (Figure 3.5 and 3.6), and NAD-ME transcript levels have been shown to cycle over the light/ dark cycle in *K. fedtschenkoi* (Dever *et al*., 2015). *PPDK* transcript levels peaked at dusk in CAM-induced *M. crystallinum* (Cushman *et al.*, 2008), and were found to be higher in the light than the dark in *Opuntia ficus-indic* (Mallona *et al.*, 2011). These findings are also consistent with the result found in scoping experiment result in Chapter 3. The malic enzymes (α- and β NAD-ME) are known to function during the day to decarboxylate malate in CAM pathway in ME-type CAM species (Cushman and Bohnert, 1997). It has not yet been demonstrated for Agave, as is the case for the majority of CAM species, whether NAD-ME or NADP-ME, or a combination of the two, is/ are used for malate decarboxylation in the light (Cushman and

Bohnert, 1997). However, *A. guadalajarana* was found to have very similar levels of NAD- and NADP-ME activity, suggesting that this Agave species may use both enzymes for malate decarboxylation in the light period (Christopher and Holtum, 1996). In the NADP-ME CAM species, *O. ficus-indica*, the *NADP-ME* transcript level showed an increase during the light period, peaking prior to dusk (Mallona *et al.*, 2011). Similarly, *M. crystallinum* is also categorised as an NADP-ME CAM species, and *NADP-ME* transcript levels peaked at the end of the light period (Cushman *et al.*, 2008).

4.2.3 Leaf Malate Concentrations in A. sisalana

A previous study by Wang and Nobel, (1998) found that the malate level in the mature leaf of *A. deserti* peaked at the end of the night and declined throughout the light period. It was therefore important to determine the level of malate in the different sections of the *A. sisalana* leaves sampled in this study in order to determine whether CAM was occurring in each leaf section. The level of malate in the white basal part of the leaf and the pale green leaf base did not vary over the light/ dark cycle (Figure 4.5). By contrast, the dark green leaf tip accumulated malate throughout the dark period, and turned the malate over in the light period, such that by 6 h into the 12 h light period the level of malate was reduced to the level detected in the leaf basal tissues (Figure 4.5). This correlated well with the fact that phase III of CAM only lasted until around the middle of the light period in similar CAM performing leaf sections from *A. sisalana* (Figure 4.1 and 4.2). These results for malate oscillations over the light/ dark cycle in the leaf tip samples also correlated well with the well-known light/dark regulation of malate levels in other CAM species. Dark CO_2 fixation leads to malic acid accumulation in the vacuole, which peaks in abundance around dawn, and light period malate decarboxylation leads to the complete turnover over the malate accumulated during the previous dark period (Black and Osmond, 2003). In the C_3-CAM tree, *Clusia rosea*, and the

inducible CAM species *M. crystallinum*, malate levels have also been reported to display the same light/ dark pattern with malate levels peaking at dawn and reaching their daily minimum in the second half of the light period (Taybi *et al.*, 2004; Häusler *et al.*, 2000) In *C. fluminensis* and *A. deserti,* which both showed a similar pattern of malate oscillation over the light/ dark cycle, reciprocating fluctuations in malate and carbohydrate, particularly sucrose levels, were reported (Wang and Nobel, 1998; Borland and Taybi, 2004).

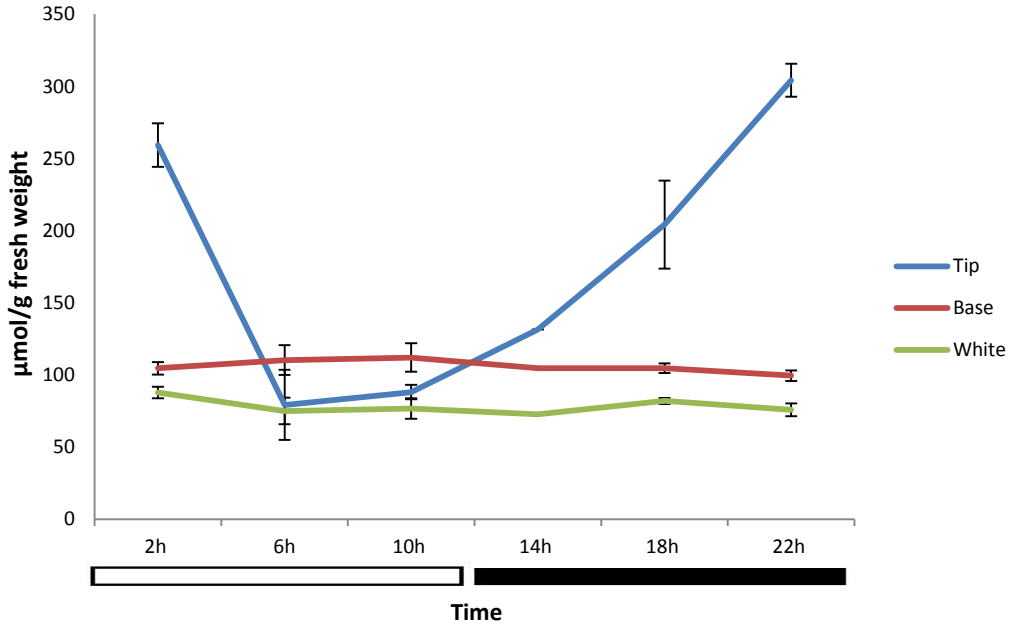


Figure 4.5 Light/ dark variation in the level of malate in the tip, base and white parts of *A. sisalana* leaves.

The concentration of malate was determined spectrophotometrically using metabolite extracts from dark green tip, pale green base, and white basal section of the youngest fully expanded *A. sisalana* leaf sampled at 4 h intervals throughout 12:12 light/dark cycles. *At 14:00, only 1 biological replicate was measured due to the loss of the other two biological replicates due to an accident.

Previous work proposed a hypothetical model for the circadian control of PEPC and CAM, involving temporal regulation in response to the daily changes in malate levels (Wilkins, 1992; Nimmo, 2000; Lüttge, 2000). A study in *K. daigremontiana* that used nitrogen gas to prevent CAM leaves from fixing atmospheric CO_2 in the dark period indicated that the circadian control of PPCK could be over-ridden by metabolite control, and also demonstrated that an increase in

the leaf malate concentration correlated with a rapid reduction in PPCK activity (Borland *et al.*, 1999). Furthermore, malate levels have been demonstrated to oscillate under LL constant, free-running conditions in CAM-induced *M. crystallinum* (Dodd *et al.*, 2003). It was therefore important to investigate the level of malate in *A. sisalana* leaves subjected LL free-running conditions in order to determine whether the circadian rhythm of CO_2 fixation recorded for the leaf tip of a mature leaf (Figure 4.2) led to a corresponding rhythm in the abundance of malate; the end product of primary CO_2 fixation in CAM species.

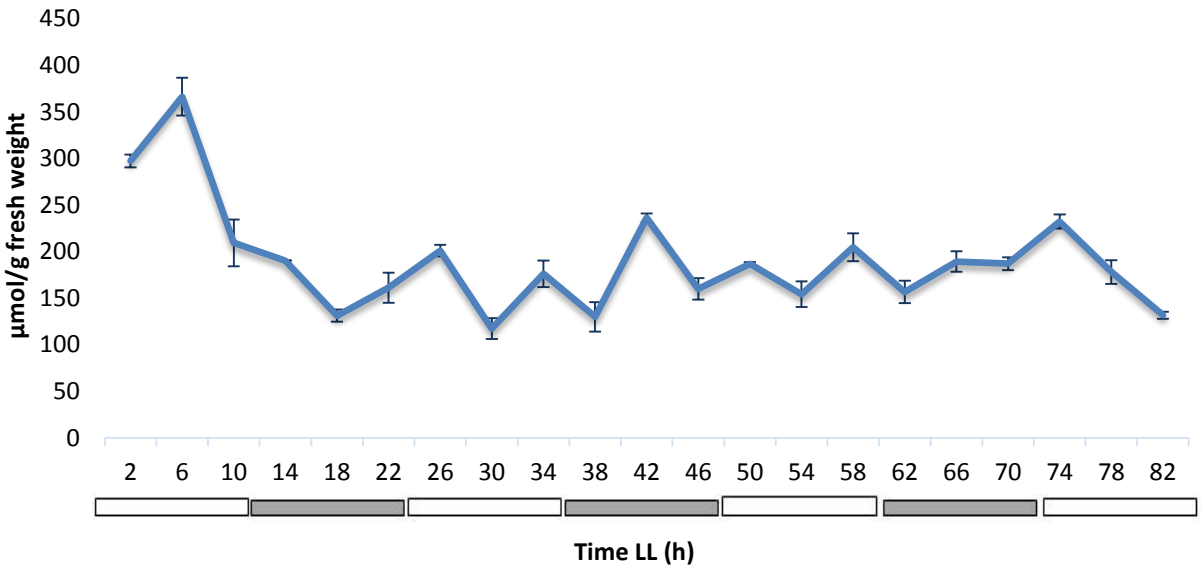


Figure 4.6 Leaf malate concentrations do not oscillate in constant light and temperature conditions in the leaf tip of *A. sisalana*.

The concentration of malate was determined spectrophotometrically using metabolite extracts from dark green tip of the youngest fully expanded *A. sisalana* leaf sampled at 4 h intervals throughout free-running constant light (LL) conditions for 82 h. The white and grey bars represent subjective light and dark respectively. Three technical replicates were used instead of biological replicates due to technical limitation of a number of similar sized plants available at the time of experiment, plus the limited room in the Snijders Microclima MC-1000 growth cabinet that could not accommodate 63 plants at the same time.

The leaf tip section of the youngest, fully-expanded leaf was sampled every 4 h throughout an 82 h time course under LL free-running conditions. Total soluble metabolites were extracted from the tip of the leaves and malate was quantified spectrophotometrically (Figure 4.6).

However, the results for malate levels in the CAM leaf tip under LL conditions showed limited, if any, evidence for a circadian oscillation in the level of malate (Figure 4.6). Buchanan-Bollig and Smith, (1984) reported a similar weak oscillation in leaf malate concentrations in CAM leaves of *K. daigremontiana.* This result contrasts strongly with the robust and large amplitude rhythm in CAM CO_2 fixation in the leaf tip (Figure 4.2), suggesting that the LL rhythm of CO_2 fixation failed to generate rhythmic malate accumulation and turnover.

4.2.4 Leaf Soluble Sugar Concentrations in A. sisalana

The source of leaf carbohydrate broken down in the dark period to fuel glycolysis and thus provide PEP for PEPC to use in CO_2 fixation varies between different CAM species (Christopher and Holtum, 1996). Some CAM species have been shown to store and use soluble hexose sugars as their carbohydrate reservoir, whereas others utilise polysaccharides such as starch. Decarboxylation of malate in the light period during CAM can be performed by either NADP-/ NAD-ME, or a combination of MDH plus PEPCK, depending on the CAM species (Christopher and Holtum, 1996). Other characteristic sets of enzymes are also found in individual CAM species according to the carbohydrate source used for PEP provision in the dark (Black *et al.*, 1996). In Agaves, sucrose is believed to be the storage carbohydrate used in glycolysis at night to provide PEP for nocturnal CO_2 fixation (Wang and Nobel, 1998).

It was thus important to determine the levels of leaf soluble sugars in the leaf soluble metabolite extracts prepared for the previously discussed malate concentration results in order to assess the potential role of sucrose, glucose and fructose levels in the daily CAM cycle in *A. sisalana*. The results for sucrose revealed that the sucrose concentration was noticeably higher in leaf tip relative to the white basal part of the leaf, and sucrose was barely detectable in the pale green base section of the leaf (Figure 4.7C). The higher concentration of sucrose in the leaf tip was consistent with the results of a previous study that reported that sucrose was

the major soluble sugar source in CAM leaves of *Ananas comosus* and *Sedum telephium* (Kenyon *et al.*, 1985). The results for sucrose levels in the *A. sisalana* leaf tip were also consistent with more recent work performed by a PhD student under supervision of Prof. Anne Borland (Dalal Albaijan, 2015, personal communication, University of Newcastle, UK) on several Agave species, none of which was *A. sisalana*. In 3 species studied, it was found that nocturnal sucrose depletion, along with CAM activity, increased from the leaf base to the leaf tip. The study also determined leaf soluble sugar levels for 14 different Agave species, and again found that CAM activity was positively correlated with nocturnal sucrose depletion. Thus, the combination of these existing results with those presented here suggests that sucrose is the most likely suspect for the provision of PEP for nocturnal CO_2 fixation in Agaves. In the leaf tip, sucrose content (3,000-8,000 µg g^{-1} fresh weight) was markedly higher than fructose and glucose (500-1,000 µg g^{-1} fresh weight; Figure 4.7C). These results agree well with those of Wang and Nobel, (1998), who reported that sucrose was the predominant sugar in all tissues of mature green (source) leaves of *A. deserti,* followed by fructose and glucose, respectively. They also found that the sucrose concentration in mature leaves of *A. deserti* exhibited a similar pattern of accumulation and depletion over the 24 h light/ dark cycle as was discovered here for the *A. sisalana* leaf tip (Figure 4.7). The leaf tip sucrose content started at a low level at dawn and increased throughout the day, peaking at the beginning of the dark period (Figure 4.7). This result is consistent with sucrose being produced as the product of photosynthesis in the light. The sucrose level then declined throughout the dark period, consistent with it being broken down and utilised for nocturnal CAM carboxylation. The fructose and glucose concentration varied relatively little over the LD cycle in most of the leaf sections, although the glucose level in the leaf tip did accumulate slightly in the light, peaking in the middle of the 12 h light period, and showed a slight light/dark regulation (Figure 4.7A and B). Glucose levels were consistently higher in the pale green leaf base and white basal leaf

tissue relative to the leaf tip, whereas fructose levels were highest in the pale green leaf base tissue and detected at a similar level in both the leaf tip and the white basal section of the leaf (Figure 4.7A and B).

In the C_3 monocot, grass species, *Lolium perenne*, fructans are synthesized and stored in the leaf base (Morvan-Bertrand *et al.*, 1999). In Agave, Wang and Nobel, (1998) reported that four fructans (neokestose, 1-kestose, nystose, and an unidentified pentofructan) were only present in the vascular tissues and phloem sap of *A. deserti* mature leaves. Fructosyltransferase activity, an enzyme that synthesises fructans, was also detected only in vascular tissues. In the photosynthetic cells, chlorenchyma (where CAM occurs), fructans and fructosyltransferases were virtually undetectable (Wang and Nobel, 1998). This suggests that fructan biosynthesis only occurs in the vascular tissues, which could be focussed on the vascular tissue in the leaf base because the base tissues function as sink tissues, receiving sucrose form the photosynthetic leaf tip that is transported via the phloem (Khan, 2001). This is consistent with the result presented here, namely fructose level was clearly higher in the leaf base (> 1,500 µg g^{-1} fresh weight) than in tip and white (< 1,000 µg g^{-1} fresh weight) (Figure 4.7B). This elevated level of fructose might be substrates for furctan synthesis in *A. sisalana* leaf base. However, the mechanisms underpinning the transport of fructans to sink tissues and/ or their synthesis and storage in sink tissue remain largely unknown.

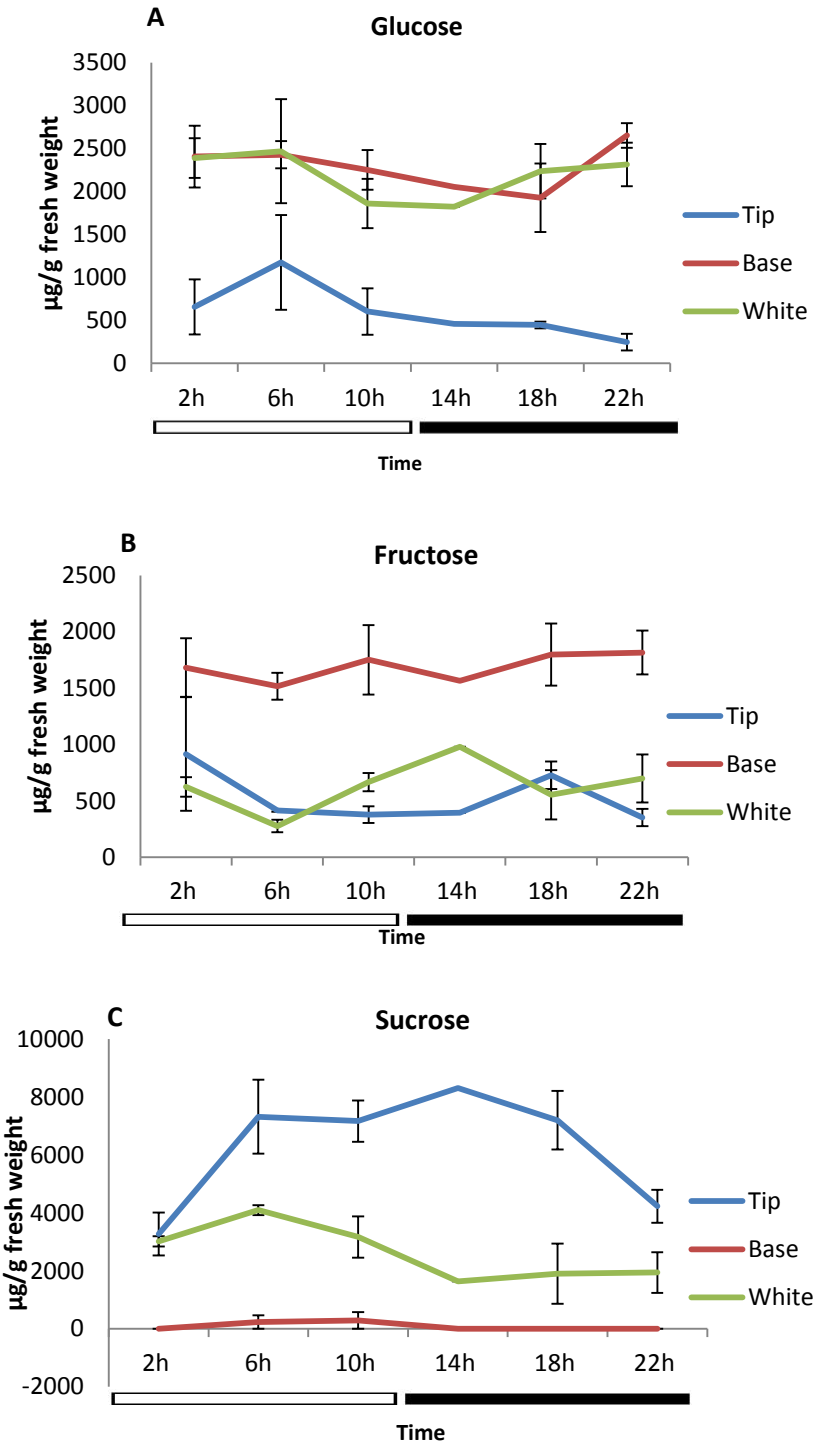


Figure 4.7 Light/ dark oscillations in the concentrations of glucose, fructose and sucrose in different sections of the youngest, fully-expanded *A. sisalana* leaf.

The concentration of soluble sugars: glucose (A), fructose (B) and sucrose (C) was measured over the light/ dark cycle for the dark green tip, pale green base, and white basal section of the youngest fully expanded *A. sisalana* leaf sampled at 4 h intervals through 12:12 light/dark cycles. *At 14:00, only 1 biological replicate was measured due to the loss of the other two biological replicates due to an accident.

One of the most logical explanations for the higher level of sucrose in the leaf tip compared to the white basal leaf tissue, and the low-to-undetectable level of sucrose in the base section of the leaf, is that sucrose was synthesised in the leaf tip due to this leaf section being the photosynthetic source tissue during the light period. The sucrose could subsequently be broken down to hexoses (fructose and glucose) due to hydrolysis by invertases (Black *et al.*, 1996; Smith and Bryce, 1992). Hexoses, especially fructose, could be transported to and stored in the leaf base and white basal tissues (proposed to be sink tissues within the leaf) (Mancilla-Margalli and Lopez, 2006; Mielenz *et al.*, 2015). Alternatively, sucrose itself may be transported in the phloem from the leaf tip to the basal sections of the leaf, and there it may be broken down to glucose and fructose. The fructose may be used for fructan biosynthesis. This explanation is supported by the higher level of fructose in leaf base relative to the other segments, and the higher level of glucose in the leaf base and white part compared to the leaf tip (Figure 4.7A). However, the glucose content in the leaf tip showed a slight light/ dark pattern of regulation. It peaked during the light period, possibly due to the production of soluble sugars via gluconeogenesis, and declined throughout the dark period, possibly due to its utilisation to provide PEP for dark CO_2 fixation (Figure 4.7A). However, the daily turnover of sucrose in the leaf tip was much greater than the turnover of glucose, and was coordinated more tightly with the predicted timing of PEP provision for PEPC activity and CO_2 fixation in the dark period, and thus the most straightforward explanation of these results would be that sucrose is used for PEP provision in the dark for CAM CO_2 fixation in the leaf tip of *A. sisalana*.

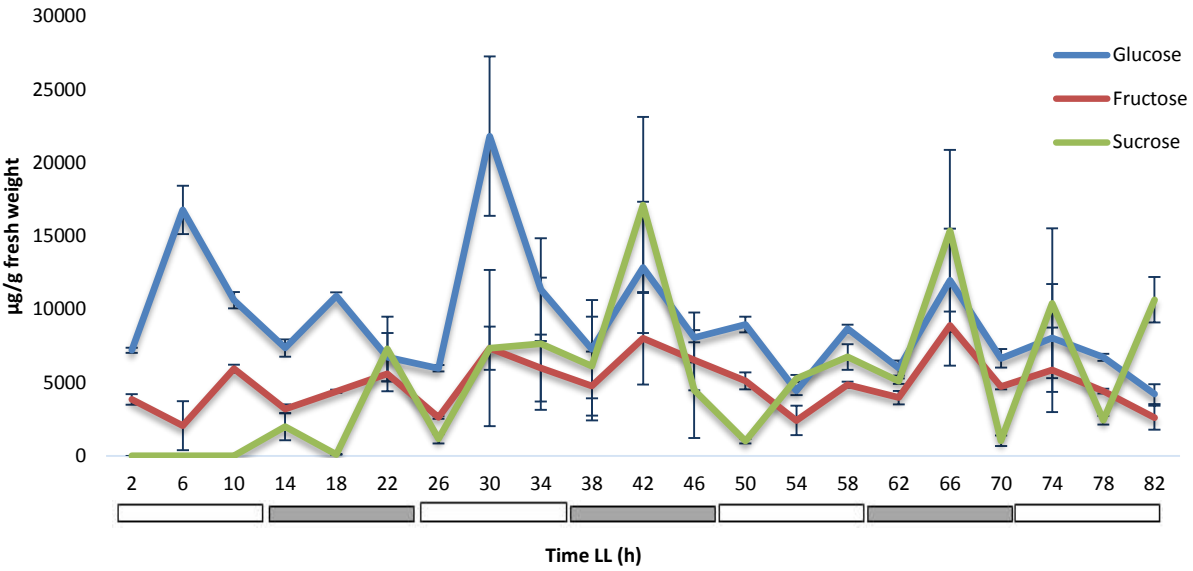


Figure 4.8 Leaf soluble sugars: glucose, fructose and sucrose concentrations demonstrate the oscillation in constant light and temperature conditions in the leaf tip of *A. sisalana*.

The concentration of glucose, fructose and sucrose was determined spectrophotometrically using metabolite extracts from dark green tip of the youngest fully expanded *A. sisalana* leaf sampled at 4 h intervals throughout free-running constant light (LL) conditions for 82 h. The white and grey bars represent subjective light and dark respectively. Three technical replicates were used instead of biological replicates due to technical limitation of a number of similar sized plants available at the time of experiment, plus the limited room in the Snijders Microclima MC-1000 growth cabinet that could not accommodate 63 plants at the same time.

The content of the 3 soluble sugars (glucose, fructose and sucrose) was also investigated using the LL leaf tip samples in order to explore the possibility of circadian clock control of the level of soluble sugars in the leaf tip sampled every 4 h throughout 82 h under LL free-running conditions (Figure 4.8). The results showed some evidence for limited oscillations in the levels of sucrose and glucose, whereas fructose did not display any evidence of an oscillation under LL (Figure 4.8). These LL results were consistent with the LD results (Figure 4.7) where sucrose in the leaf tip showed the most robust light/dark regulation, glucose showed a smaller variation over the light/ dark cycle, whereas fructose levels remained relatively stable throughout the 24 h cycle. Sucrose was undetectable at 2, 6 and 10 h after the light started for the 82 h LL experiment, but subsequently rose to a peak at the end of the first subjective dark period and then peaked in the middle of the subsequent two subjective dark periods (Figure

4.8). Sucrose levels also peaked in the dark under LD cycles, although the peak occurred at the beginning of the dark period under driven conditions (Figure 4.7). Glucose levels displayed the most robust oscillation during the first 48 h under LL conditions, with levels peaking in the middle of both the first and second subjective light period (Figure 4.8), which was consistent with the timing of the daily temporal peak of glucose in the middles of the light period in the leaf tip under LD cycles (Figure 4.7). However, glucose levels did not rise during the third subjective light period under LL, and were largely arrhythmic after the first 48 h under LL conditions (Figure 4.8).

In the leaves of the C_3 species *Arabidopsis*, starch degradation has been demonstrated to be under circadian clock control (Graf *et al.*, 2010). It has been proposed that the circadian control of nocturnal starch degradation allows for the strict temporal control of carbohydrate availability during the night, ensuring that starch is metered out until the morning (Graf *et al.*, 2010). This temporal control of starch turnover has been found to be important for the maintenance of plant productivity (Graf *et al.*, 2010). In addition, a study of the starchless *phosphoglucomutase (pgm)* mutant of *A. thaliana* revealed that sugars play an important role in modifying gene expression of approximately half of the clock-regulated genes in Arabidopsis (Haydon *et al.*, 2013). The study revealed the potential existence of a sugar-mediated entrainment of the clock. This proposal was further supported by the work of James *et al.*, (2008), which demonstrated that sugars transported from source photosynthetic tissues above-ground were able to entrain the root-specific circadian oscillator. More recently, it was found that the core central clock genes *GI*, *TOC1*, and *CCA1* were stimulated by sucrose (Knight *et al.*, 2008; Dalchau *et al.*, 2011). Mathematical and experimental studies emphasised that *GI* was associated with sucrose sensing and essential for a full response of the circadian clock to sucrose (Dalchau *et al.*, 2011). In addition, genes involved in carbohydrate metabolism are also regulated by the circadian clock (Harmer *et al.*, 2000; Rolland *et al.*, 2002; Haydon *et al.*, 2011).

These previous studies on the C_3 model species *A. thaliana* are in accordance with the results presented here, particularly the glucose and sucrose oscillations, which have been reported to have a strong association with circadian clock regulation.

4.3 Summary

The biochemical and physiological experiments described in this chapter generated a number of strongly correlated and mutually supportive datasets for the development of CAM along the length of young *A. sisalana* leaves, and for the light/dark regulation of CAM in the mature leaf tip that performs full CAM. The existing gas exchange data generated by Dr. Susanna Boxall and Dr. James Hartwell using different sections along the length of the youngest, fully-expanded *A. sisalana* leaf illustrated clearly that the leaf base performed C_3, and that older and older sections of the leaf closer to the tip performed more and more dark CO_2 fixation and less direct fixation of atmospheric CO_2 fixation in the light (Figure 4.1). The 24 h CO_2 fixation pattern in the CAM-performing leaf tip matched beautifully with the well-defined four temporal phases of CAM that span each 24 h cycle (Osmond, 1978). In addition, the gas exchange data generated for the *A. sisalana* mature leaf tip under LL free-running conditions by Dr. James Hartwell revealed that CAM CO_2 fixation in *A. sisalana* displayed a robust circadian rhythm of peaks and troughs in the level of CO_2 fixation (Figure 4.2). This LL data demonstrating a robust circadian rhythm of CO_2 fixation in the CAM-performing leaf-tip of *A. sisalana* leaves represents the first demonstration of a circadian rhythm of CAM in any monocot CAM species. In fact, robust circadian oscillations of CO_2 fixation have only been reported previously for a very small selection of dicot CAM species including *K. fedtschenkoi*, *K. daigremontiana* and *M. crystallinum* (Hartwell, 2006; Dodd *et al*., 2003).

The immune-blot analysis of the abundance of key enzymes associated with CAM revealed that the enzymes were all abundant almost exclusively in leaf tip, with only very low-to-undetectable PEPC and β-NAD-ME amount in leaf base (Figure 4.3). This suggests that CAM induction only occurred in the leaf tip. The phosphorylation state of PEPC, which in other CAM species such as *K. fedtschenkoi* has been shown to correlate perfectly with the nocturnal circadian clock mediated induction and activity of PPCK, showed strong light/ dark regulation (Figure 4.3 and 4.4). PEPC was highly phosphorylated in the dark period and dephosphorylated completely throughout the light period (Figure 4.4). The other CAM enzymes did not display strong regulation of their total abundance over the light/ dark cycle, although leaf sample amounts were insufficient to perform *in vitro* assays for the activity of each CAM enzyme, which may have provided further information regarding the light/ dark regulation of CAM enzymes in *A. sisalana*.

Malate assays over the light/ dark cycle showed that the leaf malate level oscillated in the leaf tip, peaking at the end of the dark period and reaching a daily minimum at the end of the light period (Figure 4.5), consistent with the well-defined daily oscillation of malate/ malic acid that results from dark CO_2 fixation in CAM plants (Wang and Nobel, 1998; Wang and Tobin, 1998; Black and Osmond, 2003). By contrast, the classic light/ dark oscillation in malate concentration characteristic of CAM was not detected in leaf base and white basal section of the leaf where the malate level was low and relatively constant throughout the 24 h LD cycle (Figure 4.5). However, the robust daily light/ dark oscillation in the level of malate in the CAM-performing leaf tip did not persist under LL free-running conditions (Figure 4.6). This lack of an oscillation in the level of leaf tip malate under LL conditions was at odds with the robust oscillation in CO_2 exchange by the mature leaf tip of an *A. sisalana* leaf (Figure 4.2). Further work will be required to tease apart this disconnect between the rhythm of CO_2 exchange and malate levels in the leaf, and one possible approach would be to feed ^{13}C labelled CO_2 or HCO_3^-

to the *A. sisalana* leaf tip and then to use NMR or mass spectrometry based metabolomics approaches to track the metabolic destinations of fixed $^{13}CO_2$ under LL conditions. Such experiments would reveal whether the peaks of CO_2 fixation measured under LL conditions lead to the accumulation of a distinct metabolic store of fixed carbon (e.g. sucrose or glucose rather than malate), or whether the CO_2 fixed during the peaks of CO_2 fixation that occur under LL conditions is simply respired away through futile cycling.

The measurements for the levels of the soluble sugars sucrose, glucose and fructose provided several insightful results. Firstly, sucrose was much higher in the leaf tip compared to the white basal leaf tissue, and was nearly undetectable in the pale green leaf base tissue (Figure 4.7C). Secondly, sucrose abundance displayed a very clear light/ dark cycle of abundance in the leaf tip; peaking at dusk before beginning to turnover in the first half of the dark period, presumably for PEP provision for dark CO_2 fixation (Figure 4.7C). Thirdly, fructose was found to be most abundant in the pale green leaf base section (Figure 4.7B), and it is this part of the leaf that is believed to synthesise and accumulate fructans, which are long linear and branched chains of fructose linked to a sucrose starting molecule (Morvan-Bertrand *et al.*, 1999). The fructose level did not display a strong oscillation over the light/ dark cycle consistent with the proposal that fructose does not play a role in the provision of glycolytic PEP required for CAM. Fourthly, glucose levels were highest in both the white basal leaf tissue and the pale green base leaf tissue, whereas levels were much lower in the leaf tip over the entire light dark cycle (Figure 4.7A). However, glucose levels did display an oscillation of abundance over the light/ dark cycle in the leaf tip. Finally, under LL constant free-running conditions, both sucrose and glucose were found to oscillate in abundance for at least part of the 82 h LL experiment, whereas the level of fructose was relatively constant with little if any evidence for a circadian rhythm (Figure 4.8). There are no previous reports concerning leaf soluble sugar levels measured under circadian free-running conditions in any Agave species to date. However, a

number of studies in other plant species, especially those in the C_3 model species *A. thaliana* (discussed in Section 4.2.4) show strong connection between the leaf soluble carbohydrate level and circadian clock control. Thus, the sugar level circadian oscillations demonstrated in this study support the proposal that sucrose and glucose metabolism in the CAM-performing *A. sisalana* leaf tip are under circadian clock control, and/ or that sucrose and glucose are important in the entrainment of the central circadian oscillator in *A. sisalana*, as has been demonstrated for the entrainment of the central oscillator in *A. thaliana* (Haydon *et al*., 2013).

The biochemical and physiological measurements of CAM-associated parameters presented in this chapter, together with the semi-quantitative RT-PCR scoping data from Chapter 3, confirm that CAM is only present both physiologically and metabolically in the mature, fully-developed leaf tip of the youngest fully-expanded leaf of a young *A. sisalana* plant (11-weeks after establishment of a bulbil in soil). CAM was absent in the pale green leaf base tissue, where previous gas exchange measurements indicated C_3 photosynthesis was predominating, and also absent in the non-photosynthetic white tissues found in the most basal section of the *A. sisalana* leaf. Furthermore, CAM in the tip region of *A. sisalana* leaves was found to be under strong light/dark regulation and circadian clock control. These findings clearly define the level of CAM in the different sections of the leaf used for the detailed whole transcriptome RNA-seq experiment described in chapters 5, 6 and 7.

Chapter 5

An investigation of the developmental and light/ dark regulation of CAM along the length of young *A. sisalana* leaves: comprehensive RNA-seq analysis of the genes associated with CAM

5.1 Introduction

Breakthroughs in high-throughput DNA sequencing technologies over the last 8 or 9 years have brought about a dramatic step-change in the ability of biologists to decode the genomes and transcriptomes of any living organism. This democratisation of DNA sequencing has seen genomes and transcriptomes move far beyond the tight evolutionary boundaries established by the handful of specialised model species, such as the mouse, fruit fly, Arabidopsis and yeast. Whereas previously humans were perhaps the only non-model higher eukaryotic organism for which a whole genome sequence was available, today there are at least 100 complete nuclear genomes published for higher plants alone, and many of these species would not be considered as amenable molecular genetic model species (Michael and Jackson, 2013; Michael and VanBuren, 2015).

In addition to whole genome sequencing, transcriptome sequencing, also known as RNA-sequencing or RNA-seq, has increasingly become one of the most efficient and popular approaches for decoding the expressed portion of a novel genome (Wang *et al.*, 2009; Haas and Zody, 2010). RNA-seq employs powerful next-generation sequencing technologies to

discover the presence and quantity of RNA (expressed transcripts) transcribed from a genome at a specific time, and in specific regions of tissues/ cells of a chosen study organism (Chu and Corey, 2012). RNA-seq is also known as deep sequencing due to its potential for high coverage. It sequences cDNA which is reverse transcribed in the first step of the RNA-seq library preparation. Data analysis steps within the RNA-seq pipeline usually includes transcript assembly, novel transcript discovery, and transcript quantification for differential expression analysis (Chu and Corey, 2012). To date, there are a number of next generation sequencing platforms available. Illumina, used in this study, and SOLiD provides a better sequencing depth and higher sequencing capacity compared to other platforms, allowing low abundance transcripts to be discovered (Chu and Corey, 2012). The first Illumina sequencer was released in 2006, and was quickly popular among a number of researchers due to the fact that a larger quantity of data could be obtained at a lower cost. Many researchers transitioned from using the 454 sequencing platform, which yielded sequence lengths up to 700 bp, but only around 1 million reads per run, to the Illumina platforms, because Illumina read lengths increased alongside dramatic increases in the number of reads per sequencing lane, and the technology became more cost-effective (Caporaso *et al.*, 2012; Fadrosh *et al.*, 2014; Lange *et al.*, 2014). At present, the Illumina MiSeq platform can generate the longest paired-end reads (300 bp) compared to the other Illumina platforms. However, the Illumina HiSeq-2500 system (used in this study) was the platform with the greatest overall output at the time that the sequencing was submitted to the University of Liverpool CGR as it could generate 4 billion paired-end fragments with up to 125 bp read length for each read in a single run (Hodkinson and Grice, 2015).

Transcriptome sequencing (RNA-seq) has the advantage that it avoids the need to sequence through often large and repetitive non-coding regions of complex plant genomes and instead focuses specifically on the transcripts that are present in specific tissues and organs, and/ or

experimental treatments of interest. In addition, RNA-seq not only facilitates the assembly of transcripts representing all of the genes being expressed in a particular tissue or pool of tissues, but can also permit the quantification of the abundance of a particular transcript in that tissue. Abundance measurements are determined through counting the number of sequence reads that map to each assembled transcript; the number of reads mapping to each transcript, or gene if a complete genome sequence is available, is directly proportional to the level of RNA that was present for that gene in the original RNA sample. Quantitative RNA-seq values are often expressed as either "Fragments per Kb of contig length per million reads" (FPKM), "reads per Kb of contig length per million reads" (RPKM), or "transcripts per million reads" (TPM). FPKM and RPKM are interchangeable; FPKM is often used with data from the Illumina sequencing system, as the method of sequencing generates a pair of short sequence reads (generally between 100 and 150 bp) from both ends of a short fragment (generally around 300 – 400 bp). The original 300 – 400 bp fragment of cDNA is thus the "fragment" in the calculated FPKM values.

Today, RNA-seq has been used and reported widely as a rapid and powerful technique that facilitates the discovery of novel genes of interest in previously unsequenced species. Even in model species such as *A. thaliana,* that were previously thought to have comprehensive genome and transcriptome resources, results from one of the first plant based RNA-Seq studies revealed a huge number of novel transcribed regions that had remained unexplored in *A. thaliana* until the advent of high-throughput sequencing technologies (Weber *et al.*, 2007; Lister *et al.*, 2008). RNA-seq studies have been used for the *de novo* assembly and quantitative analysis of the transcriptomes for a myriad of non-model plant species and crops.

A traditional Expressed Sequence Tag (EST) cDNA sequencing project has been reported for normalised cDNA generated from mixed tissues of *A. sisalana* (Zhou *et al.*, 2012). The authors

identified 3320 unigenes from the random selection and sequencing of 3875 clones from a cDNA library using traditional Sanger dideoxy sequencing of each cDNA clone (Zhou *et al*., 2012). By contrast, a much more comprehensive analysis of the transcriptome of *A. deserti* and *A. tequilana* was achieved using RNA-seq with Illumina high-throughput sequencing technology (Gross *et al*., 2013). RNA samples were generated from leaves, roots, stem and juvenile tissues for *A. tequilana*, and from roots, folded leaves and meristem, four different sections of the leaf (evenly spaced in length from leaf base to leaf tip), ramets, and the proximal leaf for *A. deserti*. Individual sequencing libraries were made from each tissue type sampled for both *A. tequilana* and *A. deserti*, and in total, 293.5 Gbp of Illumina sequence was generated for *A. tequilana* and 184.7 Gbp of sequence was generated for *A. deserti*. The reads were assembled using Rnnotator, which resulted in just over 204,000 contigs for *A. tequilana* and almost 129,000 contigs for *A. deserti*. Both assemblies had an average length for the assembled contigs above 1300 bp and both assemblies were predicted to encode approximately 35,000 protein-coding genes (Gross *et al.*, 2013). The authors concluded that their *A. tequilana* and *A. deserti* transcriptome assemblies were almost complete in terms of coverage.

The work of Gross *et al*. (2013) included a study of the changes in gene transcript levels along the proximal-distal axis of two fully developed *A. deserti* leaves sampled from young plants (leaves approximately 12 cm in length). This RNA-seq study in *A. deserti* was very similar to the study of the developmental regulation of the *A. sisalana* proximal-distal (base-tip) leaf transcriptome presented in this thesis. However, Gross *et al*. (2013) did not use biological triplicate samples, and they did not investigate light and dark samples and thus capture the dynamic light/ dark regulation of the transcripts in the different aged sections of the *A. deserti* leaves. In fact, the description of the timing of the leaf sampling was rather vague "*A. deserti* tissues were collected from well-watered plants near mid-day"; no details were provided as to

the length of the light and dark cycles under which the plants were grown (Gross *et al*., 2013). It should be noted that all of the *A. sisalana* leaf sampling and Illumina RNA-sequencing work described in this thesis had been completed before the publication of the report by Gross *et al.* (2013).

The proximal-distal analysis of the developmental regulation of the leaf transcriptome in *A. deserti* revealed that genes encoding proteins associated with both photosynthetic electron transport, the Calvin cycle and the CAM adaptation of photosynthetic CO_2 fixation increased in transcript abundance in the distal (tip) portion of the leaf relative to the proximal (basal) section of the leaf (Gross *et al*., 2013). However, in terms of CAM pathway genes, they only reported on the identification of assembled contigs annotated as *PPC*, *MDH* and *PPDK,* and thus did not provide a comprehensive analysis of candidate CAM-associated genes for *A. deserti,* or their regulation along the proximal-distal leaf developmental profile. In fact, only in the case of *PPC* did they suggest that they had identified the 'photosynthetic' member of the *PPC* gene family based on its high RPKM value in the leaf tip relative to the leaf base (Gross *et al*., 2013). Furthermore, the lack of RNA-seq data for dark samples of the leaf tip prevented them from identifying the gene(s) encoding the CAM-associated *PPCK*s, and they did not make any attempt to investigate the genes associated with sucrose and fructan metabolism, which are the candidate carbohydrate sources used for PEP provision for primary CO_2 fixation by PEPC in the dark period in Agave. Overall, whilst the study of Gross *et al.* (2013) achieved a very high quality *de novo* assembly of the transcriptome of *A. tequilana* and *A. deserti*, the lack of detailed quantitative analysis of the genes that were induced in the leaf tip relative to the leaf base meant that the study only provided a very preliminary and limited insight into the genes that are induced concomitant with CAM in the leaf tip in these Agave species.

In a separate study completed prior to the start of this PhD by the Hartwell lab (University of Liverpool), Roche 454 pyrosequencing was used to generate a preliminary assembly of the *A. sisalana* leaf transcriptome. Light and dark samples of leaf tip full length cDNA were sequenced using the 454 system, and generated 1.2 million 454 reads with average read length of 307 bp. This provided 368 Mbp of transcriptome sequence data, which were assembled using MIRA, yielding 90,044 contigs (Boxall, Gregory and Hartwell, unpublished data). Key CAM genes such as *AsPPC, AsPPDK, AsPPCK, AsV-ATPas*e subunits, malate dehydrogenase, *AsNAD*- and *AsNADP-ME*, and various glycolytic enzymes were identified along with Rubisco, PSB (photosystem II subunits) and *AsCAB* genes as the most abundant transcripts in the *A. sisalana* leaf tip transcriptome. This preliminary assembly of the *A. sisalana* transcriptome provided an invaluable sequence database for the RT-PCR primer design required for the study of specific CAM, circadian clock and sugar metabolism associated genes that was described in Chapter 3.

The results of the preliminary semi-quantitative RT-PCR experiments presented in Chapter 3, combined with the biochemical and physiological analysis of the *A. sisalana* proximal-distal leaf developmental gradient and its temporal light/dark regulation presented in Chapter 4, provide a robust framework underpinning the selection of the leaf segments and time points for the Illumina RNA-seq experiment that provided the major dataset presented in this thesis. Due to budget limitations, only the white basal, pale green basal, and dark green tip part of the youngest fully expanded leaf of 11-week-old *A. sisalana* sampled at 10:00 light (2 h before dusk) and 22:00 dark (2 h before dawn), were sequenced using 3 biological replicates, leading to a total of 18 total RNA samples that were submitted to the Centre for Genomic Research at the University of Liverpool.

This chapter presents the comprehensive analysis of the genes involved in CAM and the circadian clock using data obtained from RNA-seq reads generated using the Illumina HiSeq-2500 sequencer. The RNA-seq analysis pipeline proceeded through a number of key steps including RNA preparation and quality control (QC), library preparation and sequencing, *de novo* transcriptome assembly, annotation, and the subsequent quantitative and statistical analysis that allowed the identification of differentially expressed genes and thus candidate transcription factors that may play a role in the light/ dark and circadian clock control of CAM in *A. sisalana*. The aim of this analysis was to obtain a transcriptome-wide and comprehensive insight into CAM-related genes in *A. sisalana*. The identification of novel putative CAM genes and provision of this large-scale transcriptome data resource for the important Agave crop species *A. sisalana* will facilitate future studies of the biochemistry and functional genomics of CAM in *A. sisalana*.

5.2 Result and discussion

5.2.1 RNA-sequencing output and read quality

The bar-coded libraries were multiplexed across 3 lanes of the Illumina HiSeq-2500 with 6 libraries per lane. The sequencing run worked well with a minimum of approximately 50 million reads recovered per library/ RNA sample, and a consistent read length for the R1 sense strand and R2 antisense strand paired-end reads of 100 bp (Supplementary Figure 5.4). Overall, the 18 libraries generated approximately 90 Gbp of RNA-seq data from all of the different leaf sections and time points (Supplementary Figure 5.3 and 5.4). Whilst this covered fewer total base pairs than the published Illumina RNA-seq data generated for *A. tequilana* (293.5 Gbp) and *A. deserti* (184.7 Gbp) (Gross *et al.,* 2013), it provided a much greater read depth coverage than the *A. sisalana* 454 transcriptome sequence dataset which consisted of

1.2 million 454 reads with an average read length of 307 bp spanning a total of 368 Mbp of data (Boxall, Gregory and Hartwell, unpublished data).

5.2.2 De novo assembly and annotation of the Illumina RNA-seq data

***De novo* assembly**

A *de novo* assembly of the Illumina sequencing reads was performed using the assembly pipeline called Trinity in order to generate a reference transcriptome for the *A. sisalana* leaf (Grabherr *et al.,* 2011). In comparison with the other *de novo* transcriptome assemblers, Trinity generally obtains more full-length transcripts spanning a wide range of transcript abundance levels in the original RNA samples. Its sensitivity is comparable with genome-alignment approaches. Trinity applies 3 separate software packages, named Inchworm, Chrysalis, and Butterfly (described in Section 2.10.1), to assemble the reads (Grabherr *et al.*, 2011). The paired-end, strand-specific forward (R1) and reverse (R2) reads for all 18 samples were assembled. The singlet (R0) reads were excluded as they contained sequences whose pairs had been removed due to poor sequence quality, or adapter contamination. The R0 reads were however only recovered in small numbers and this made no difference to the assembly. Trinity joins assembled transcripts together into clusters according to their shared sequence content. This cluster of shared-content transcripts is roughly defined as a “gene”. Trinity enables an identification of paralogous genes and alternatively spliced transcript isoforms that diverged from recently duplicated genes (Grabherr *et al.*, 2011). Consequently, distinct paralogous genes and alternatively spliced transcript isoforms were identified and reported, generating a very large number of total assembled Trinity genes. This assembly includes paralogues, with each gene being represented by one or more transcript isoforms. The assembly presented here spanned a very high number of Trinity genes (671,886), and

spliced transcript isoforms (941,989) (Table 5.1). These numbers are considerably higher than those previously found in the previously published RNA-seq study for two Agave species published by Gross *et al.*, (2013). This report generated 139,525 loci and 204,530 contigs for *A. tequilana*, and 88,718 loci and 128,869 contigs for *A. deserti* using the Rnnotator assembler. These contigs were assembled from a larger total amount of sequence data for *A. tequilana* (293.5 Gbp) and *A. deserti (*184.7 Gbp). The algorithm that Trinity uses to differentiate and report paralogous and spliced transcripts leads to it assembling and retaining a greater number of these transcript forms (Grabherr *et al*., 2011). The number of *A. sisalana* contigs in the new Trinity assembly reported here are also higher than the number assembled through the previous *A. sisalana* 454-transcriptome assembly generated in the Hartwell lab, which included a total of 90,044 contigs. This assembly was generated with the MIRA assembler and was assembled from a much smaller total set of transcriptome sequence data (368 Mbp).

Protein-coding sequence identification

Protein-coding sequences were identified using TransDecoder, a tool that enables the identification of likely protein-coding regions in the newly assembled transcript sequences. It identifies candidate protein-coding regions based on composition of the nucleotide and the length of the longest open reading frame (ORF) (Haas *et al.*, 2013). TransDecoder generated 118,807 protein-coding sequences that integrated the overlapped predicted protein sequences from the 941,989 assembled isoforms (Table 5.1). In line with the higher number of contigs assembled, the number of protein-coding sequences reported here is again greater than number of proteins reported by Gross *et al.*, (2013), which identified 34,870 protein-coding sequences for *A. tequilana* and 35,086 for *A. deserti*.

Assembly evaluation

The N50 value, median contig length, average contig length, total assembled bases and other statistics were calculated based on both all transcript contigs and only longest isoforms per gene, using Quast and TrinityStats.pl (Table 5.1). The total number of bases assembled (505.3 Mbp) was greater than the values reported previously for *A. tequilana* (204.9 Mbp) and *A. deserti* (125.0 Mbp) (Gross *et al.*, 2013). Contig N50 is a weighted median value whereby 50 % of the contigs in the assembly are longer than the N50 length (Miller *et al.*, 2010). The N50 value for the Trinity assembly of *A. sisalana* reported here was 689 bp for the assembly based on all transcript contigs and 625 bp for the assembly based on only longest isoform per gene (Table 5.1). The N50 value of this assembly (a combination of all 18 samples), based on all transcript contigs, is lower (689 < 979 bp) than the 1-biological replicate (6 samples) assembly that was previously performed for comparison and optimisation purpose. In contrast, it was higher (625 > 555 bp) when based on only longest isoform per gene. However, the N50 value is statistically primitive and could often be misleading when compared with other assemblies of different sizes (Li *et al.*, 2014) due to the fact that it only assesses the continuity but not accuracy of contigs (Salzberg *et al.*, 2012). Thus, a more qualitative evaluation method was needed. In addition, the Trinity assembly also generally contained longer and more complete sequences for known CAM genes when compared to the MIRA 454-transcriptome assembly generated previously in the Hartwell lab by Richard Gregory.

Table 5.1 Statistics of Trinity assembly generated using a combination of Quast, Trinity built-in tool "TrinityStats.pl" and TransDecoder

General statistics	
Trinity 'genes'	671,886
No. of Trinity 'transcripts (spliced isoforms)'	941,989
No. of Trinity 'transcripts' (> or = 1000 bp)	109,662
No. of Protein-coding sequences	118,807
Largest contig	18,632 bp
Percentage of GC	39.03
Statistics based on all transcript contigs	
Contig N50 length	689 bp
Median contig length	352 bp
Average contig length	536.43 bp
Total assembled bases	505,308,127 bp
Statistics based on only the longest isoform per gene	
Contig N50 length	625 bp
Median contig length	333 bp
Average contig length	507.19 bp
Total assembled bases	340,772,327 bp

The assembly was further evaluated using CEGMA to assess the completeness of the assembled transcriptome. CEGMA uses various mapping tools to align protein sequences from the predicted Core Eukaryotic Genes (CEGs), a data set of 458 core proteins that are highly conserved in a broad range of eukaryotic species, to the predicted proteins within the assembled transcriptome. It then reports the proteins that mapped to the transcriptome tested, which are thus identified as present in the transcriptome assembly (Parra *et al.*, 2007). Almost all of the CEG proteins were found within the Trinity assembly of the *A. sisalana* transcriptome (n = 454 out of 458 CEGs; 99.13 %). This result demonstrated that the Trinity assembly had a very high level of completeness for this core eukaryotic set of genes and their

encoded proteins. Furthermore, the CEGMA output also generated a table containing a summary of which of the subset (Group) of the 248 most highly-conserved CEGs are present either partially or completely in the transcriptome (Parra *et al.*, 2009).

Table 5.2 Statistics of the completeness of the transcriptome based on 248 ultra-conserved CEGs

'Complete' indicates the predicted proteins from the set of 248 CEGs that aligned 70% or higher of the target protein in the transcriptome. 'Partial' represents the CEGs with less than 70% aligned protein length but still exceeds a pre-computed minimum alignment score. A protein estimated to be 'Complete' was also included in the 'Partial'. Group 1 contains the least conserved of the CEGs with Group 4 containing the most highly conserved CEGs. Prots indicates the number of 248 the ultra-conserved CEGs present in transcriptome. Completeness represents a percentage of 248 ultra-conserved CEGs present. Total is a total number of CEGs present including putative orthologs. Average is an average number of orthologs per CEG. Ortho is a percentage of detected CEGS that contain more than 1 ortholog.

	Prots	Completeness (%)	Total	Average	Ortho (%)
Complete	**236**	**95.16**	**851**	**3.61**	**89.83**
Group 1	62	93.94	218	3.52	95.16
Group 2	53	94.64	208	3.92	90.57
Group 3	57	93.44	191	3.35	85.96
Group 4	64	98.46	234	3.66	87.5
Partial	**246**	**99.19**	**1047**	**4.26**	**98.37**
Group 1	66	100	277	4.2	100
Group 2	56	100	247	4.41	100
Group 3	59	96.72	249	4.22	98.31
Group 4	65	100	274	4.22	95.38

The result in Table 5.2 shows that 95.16 % (236 of 248) of the ultra-conserved CEGs are completely represented in the transcriptome. The most conserved CEGs, group 4, shows the highest percentage (98.46 %) of completeness relative to the other less highly conserved groups (Table 5.2). Furthermore, 99.19 % (246 of 248) of the ultra-conserved CEGs were estimated by CEGMA to be at least partially represented in the Trinity transcriptome assembly. The most conserved CEGs, group 4, were covered with 100 % completeness, along with the less highly conserved CEGs in group 1 and 2. However, 96.72 % of the group 3 CEGs were

represented at least partially in the Trinity assembly. These findings considered in combination with the 454 out of 458 (99.13 %) mapped proteins from the total CEGs set, confirmed that the transcriptome assembly covered a very high number of the CEGs, suggesting a high quality and representative assembly was achieved. Hence, the *de novo* assembly generated in this study was likely to provide a comprehensive coverage of the transcriptome of *A. sisalana,* thus supporting its further use for downstream quantitative analysis of the regulation of the leaf transcriptome along the proximal-distal axis of the youngest fully-expanded leaf.

Annotation

The transcriptome was annotated using Trinotate, a comprehensive annotation tool included in the Trinity package (Grabherr *et al.*, 2011), and designed for functional annotation of *de novo* assembled transcriptomes. Trinotate exploits several different well-known functional annotation approaches. It also uses the most likely longest-ORF peptide candidates (protein-coding sequences) previously generated from the assembly using TransDecoder as query protein sequences for BLASTP homology searching of public databases of annotated and characterised proteins and protein domains and motifs. All of the functional annotation data derived from all transcripts is then merged into a database called SQLite. Ultimately, all individual trasncripts were annotated with annotation information derived from different databases including PFAM domain (Punta *et al.*, 2012), UniProt (Magrane and Consortium, 2011), and Gene Ontology (GO) (Ashburner *et al.*, 2000). A master annotation report for the whole assembly was generated. This facilitated the identification and characterisation of novel genes of interest in the next stage of the analysis.

5.2.3 Differential expression analysis

Transcript abundance (expression level) was estimated using a built-in Trinity tool called RSEM. RSEM was chosen as it allows accurate transcript quantification in species without available reference genomes but where a *de novo* transcriptome assembly is available, as was the case here. RSEM has been reported to achieve a higher or comparable quantification accuracy to other similar tools (Li and Dewey, 2011). A major complexity of quantification is that RNA-Seq reads often map uniquely to more than one gene or spliced transcript isoform. To solve this problem, RSEM utilises an expectation-maximization algorithm to estimate maximum likelihood expression levels. RSEM estimates the number of fragments derived from a given gene or isoform. This count is commonly the expected number of unfiltered and alignable fragments obtained from a gene or isoform given the maximum likelihood abundances. Thus, these counts can be used for a differential expression analysis using method such as edgeR (Li and Dewey, 2011). In this study, the original raw reads from the Illumina sequencing of each of the 18 libraries generated from the replicated total RNA samples were mapped back to the Trinity assembled transcripts using Bowtie2, which is included in the RSEM pipeline. As output, RSEM reported the abundance estimation of reads per Trinity 'gene' and per spliced transcript isoform. The expected read counts estimated were consequently used in the differential expression analysis using edgeR in the next analysis stage. In addition, FPKM values were also generated allowing the ability to compare the expression level of transcripts across different libraries (samples). The mapping percentage of each sample was estimated using 'samtools flagstat' from an RSEM alignment output (bam file). Tip samples from both light and dark time points exhibited a higher percentage of properly paired mapping relative to the pale green base and white samples (Figure 5.1). The properly paired reads mapping implies that both

paired-end reads are mapped to the same transcript, oriented towards each other with a sensible insert size (Li and Dewey, 2011).

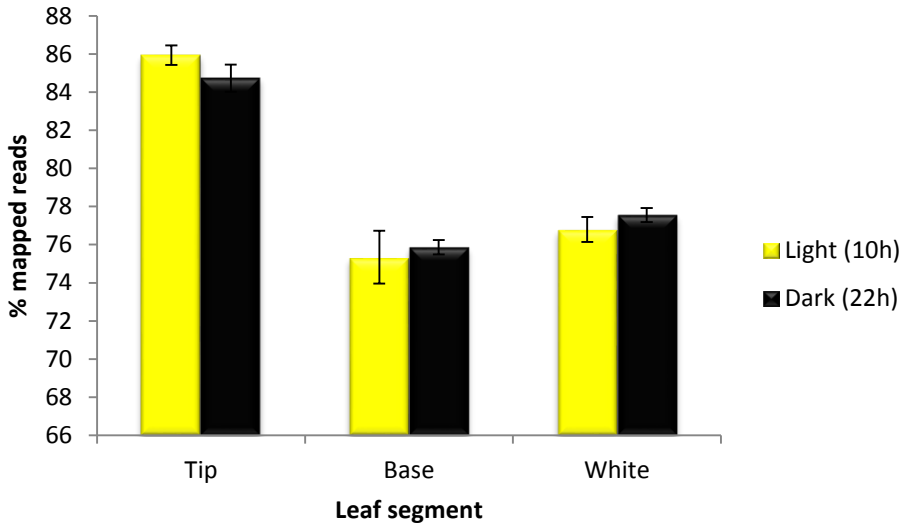


Figure 5.1 Percentage of mapped reads into contigs of each sample using RSEM

The percentage of reads that mapped to assembled transcripts was estimated using 'samtools flagstat' from alignment output file (bam). The reads were derived from RNA-seq data of the youngest fully expanded *A. sisalana* leaf tip, base and white segments sampled at 10:00 (light) and 22:00 (dark) with 3 biological replicates per sample.

The differential expression analysis was performed by Dr. Yongxiang Fang (a bioiformatician in the Centre for Genomic Research, University of Liverpool) using edgeR to estimate gene transcript abundance at the Trinity 'gene' level (Robinson *et al.*, 2010). For the quantitative analysis, 671,886 Trinity 'genes', each containing one or more alternatively spliced isoforms, were analysed in terms of the regulation of their transcript abundance both spatially and temporally between different leaf sections sampled in the light (10:00) and dark (22:00). Differential expression was investigated at the gene level rather than gene isoform level as the goal was to achieve a transcriptome-wide view of the regulation of each gene of interest. The total transcript abundance per gene can potentially provide a superior and wider view of the transcriptome, improving the success rate for the identification of novel candidate genes of interest. However, it should be noted that further analysis at the isoform level could be

performed in order to investigate the regulation of specific genes in greater detail; in particular isoform level analysis can provide information about the regulation of a gene at the level of its alternative splice forms.

Despite the fact that alternative splicing is well known to play an important role in gene regulation and proteome divergence in animals, its importance in plants has not yet been investigated in great detail (Potenza *et al.*, 2015). Alternative splicing in plants and its role in the regulation of plant gene expression remains a relatively unexplored topic (Reddy *et al.*, 2013). However, since the next generation sequencing techniques have been introduced, the number of plant genes supposed to be under alternatively splicing events has increased dramatically (Potenza *et al.*, 2015). To date, the analysis of plant transcriptomes based on high-throughput sequencing has proposed that up to approximately 60% of pre-mRNAs from intron-containing genes are alternatively spliced producing a large number of mRNA isoforms (Filichkin *et al.*, 2015). Some recent studies have revealed the importance of alternative splicing in the regulation of plant metabolism (Reddy *et al.*, 2013) and the response of the central circadian clock to temperature (James *et al.*, 2012). These studies have also discovered that the regulation of alternative splicing in plants can be impacted by cell type, developmental stage, and the environment (Filichkin *et al.*, 2015). It has been reported that a few plant genes with alternatively spliced isoforms have been shown to have an *in vivo* functional impact in a wide range of physiological and developmental processes, including the involvement of alternative splicing in the regulation of photosynthesis and starch metabolism (Carvalho *et al.*, 2013). These findings emphasise that it will be valuable in future to reanalyse this RNA-seq dataset for *A. sisalana* leaf development and its light/ dark regulation in order to dissect out the role of alternative splicing in the regulation of genes required for CAM and its regulation in response to the light/ dark cycle and the circadian clock.

The differential expression analysis compared the transcript abundance of all ~672,000 assembled *A. sisalana* Trinity 'genes' between the dark green leaf tip, pale green leaf base, and white basal tissue at the most basal/ proximal portion of the leaf, sampled both in the light (10:00; 2 h before dusk) and the dark (22:00; 2 h before dawn). The samples from the different leaf segments and time points were organised into 6 groups, each with 3 biological replicates, and these were compared to one another using edgeR analysis of the mapping frequency data for each gene generated by Bowtie2. In addition, the total number of transcripts mapping to each Trinity 'gene' for the tip, base and white samples, combining the light and dark values, were also compared with each other, making a total of 12 contrasts that were tested for the differential regulation of genes (the contrasts compared are described in detail in Section 2.10.3).

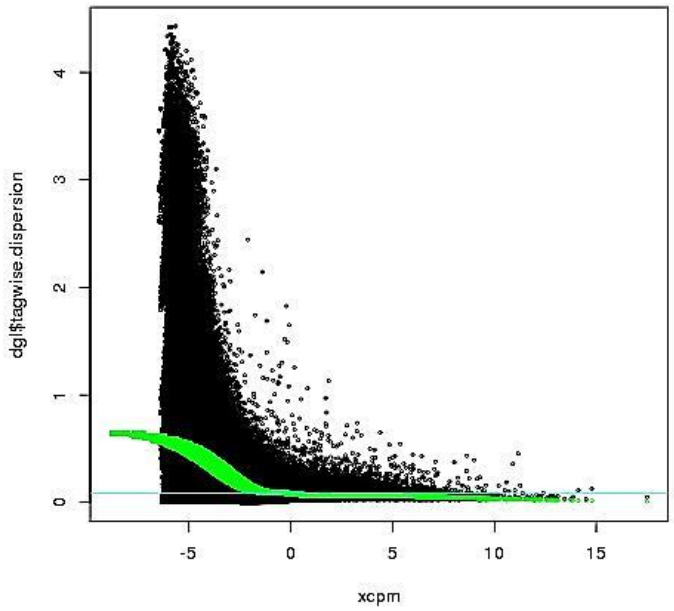


Figure 5.2 Dispersion plot of all transcripts and samples.

Dispersion plot was generated to illustrate variability of data among transcripts. The thin light blue line indicates the common (average) dispersion across all samples, irrespective of transcripts. The thick green line models the dispersion by transcript abundance (trend). The block dots represent the estimates of dispersion for each transcript. Y-axis indicates tagwise dispersion values. X-axis represents LogCPM values. The plot was generated by Dr. Yongxiang Fang, CGR, University of Liverpool.

A dispersion plot was generated to investigate the variability of data among all of the transcripts, especially when biological replicates were used (Figure 5.2). Dispersion values were higher for rare transcripts, suggesting that these rare transcripts may be responsible for high levels of biological variation even though their counts per million (CPM) values were low. The graph reveals that the majority of transcripts were rare transcripts (Figure 5.2). This shows that the ability to detect the differential expression of rare transcripts is not limited by sequencing depth but more likely due to natural variation amongst the biological replicates (Robinson *et al.*, 2010). In edgeR analysis, tagwise (genewise) dispersion values were estimated for individual transcripts (Trinity 'genes') using count data previously generated by RSEM. This allows the estimation of biological variability among samples. The tagwise dispersion values were normalized and used to fit the negative binomial model. edgeR employs an empirical Bayes method to moderate dispersion values across transcripts, squeezing them towards a common value. Differential expression of transcripts was then analysed and statistics were calculated. The CPM was used by edgeR to reflect expression levels (Robinson *et al.*, 2010). This analysis resulted in the identification of 141,366 differentially expressed (DE) genes from all of the 12 contrasts that were analysed. A master table was generated, which represents the unified set of the DE genes from all 12 contrasts and contains all of the DE genes and their statistics (data not shown). In other words, each gene contained in the table was differentially expressed under at least in one of the contrasts that were analysed.

To obtain a broad view of the DE analysis, the principal components were plotted to assess the variation in transcripts among and between the different groups of samples (Figure 5.3). The first principal component, containing the most highly variable values for the samples, was plotted against the second principal component, which captured the second greatest variance (Figure 5.3A). In this plot, the tip samples, both in the light and dark, were very clearly

separated from leaf base and white basal samples, which tended to be more closely correlated with one another (Figure 5.3A). This first and second component PCA plot also showed a clear separation between the tip light (in black) and dark samples (in dark blue). When the second component was plotted against the third component, which contains the least variance, both the tip and basal samples were clearly separated, as were the pale green base and white base samples (Figure 5.3B). However, the light and dark samples from each leaf section were placed closer together, although most of the biological replicates were still clearly separated from one another (i.e. W_L_2 and W_L_3 are clearly separated from W_D_2 and W_D_3 (Figure 5.3B)).

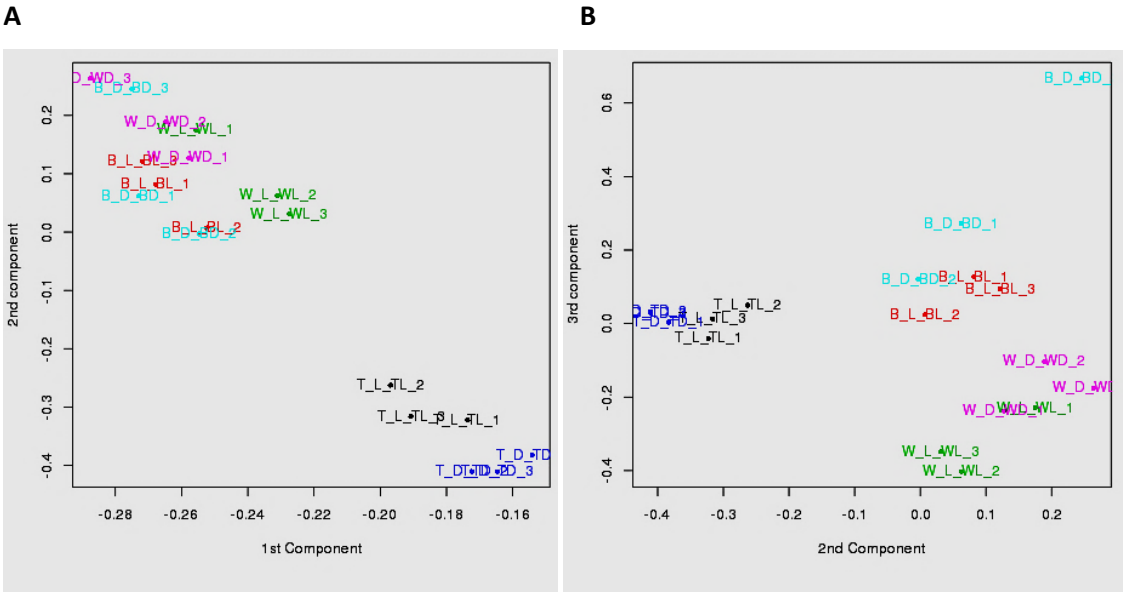


Figure 5.3 Principal component analysis plotted to visualise variation among and between the groups of RNA-seq samples.

Principal component analysis (PCA) plots were generated to visualise the variation of groups of samples. The 1st principal component was plotted against the 2nd component (A) and the 2nd component against the 3rd component (B). (A) reveals the greatest variance among groups of samples. (B) exhibits the remaining variance. The abbreviations of the youngest fully expanded *A. sisalana* leaf samples indicate as follows; TL = Tip Light (10:00; black), TD = Tip dark (22:00; dark blue), BL = Base Light (red), BD = Base Dark (light blue), WL = White Light (green), WD = White Dark (pink). The number at the end of each sample indicates the biological replicate. The PCA plot was generated by Dr. Yongxiang Fang, CGR, University of Liverpool.

To visualise the similarity and correlation of transcriptomic profiles between replicates and samples, a heatmap was generated displaying the Pearson correlation coefficient of transcript abundance across the 3 biological replicates (Figure 5.4). Overall, the biological replicates of each sample were most similar to the other biological replicates within each group of samples, indicated by the more yellow, orange, red and brown colours representing higher Pearson correlation coefficients (Figure 5.4). Both the light and dark leaf tip samples (TL and TD) showed relatively low levels of correlation with the base and white samples, indicated by the dark blue colour which represents Pearson correlation coefficient around or below 0.7. The correlation was lower between the tip and white samples than it was between the tip and base samples. In terms of light/dark correlations, it was most striking that the tip dark (TD) samples showed the lowest correlation with both the light and dark white basal samples (WL and WD), while the correlation increased slightly when comparing the tip dark (TD) and pale green base samples from both the light and dark (BL and BD). Tip light (TL) samples showed a similar level of correlation with the tip dark (TD), and both leaf base samples (BL and BD). However, the TL correlation with the WL and WD samples was substantially lower (Figure 5.4). The base and white samples were correlated relatively strongly with one another (yellow and turquoise colours) relative to their level of correlation with the tip samples (blue colours). These results indicate that the correlation between the level of transcripts for each gene in the CAM performing leaf tip, especially during the dark period, and the level of transcripts from both the C_3–performing leaf base, and the non-photosynthetic white basal leaf segment was lower than the correlation between transcripts from base and white samples (Figure 5.4). This supports and emphasises the validity of the interpretation of the principal component analysis (Figure 5.3).

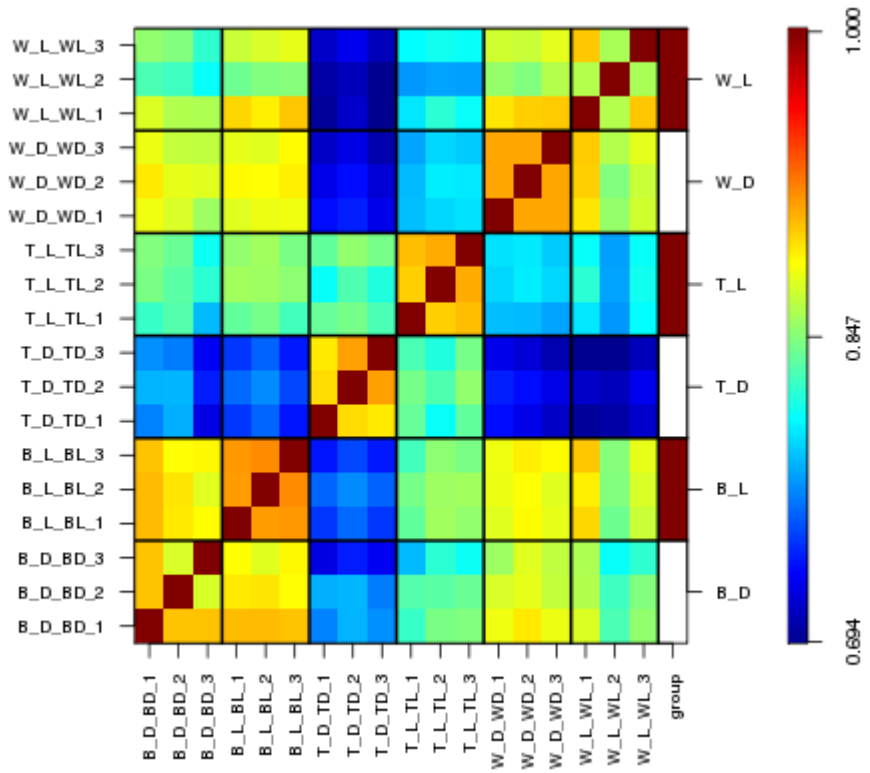


Figure 5.4 Pearson correlation coefficient heatmap displaying agglomerative hierarchical clustering of samples

A Pearson correlation coefficient heatmap of the transcript abundance across the 3 biological replicates of the samples was generated. The different colours represent different values as indicated in the multi-coloured bar on the right with a range of values. The abbreviations of the names of the youngest fully expanded *A. sisalana* leaf samples are as follows; TL = Tip Light (10:00), TD = Tip dark (22:00), BL = Base Light, BD = Base Dark, WL = White Light, WD = White Dark. The number at the end of each sample indicates the biological replicate. The heatmap was generated by Dr. Yongxiang Fang, CGR, University of Liverpool.

5.2.4 The identification of novel differentially expressed genes of interest and candidate CAM-associated genes

The identification of CAM genes

The DE genes were grouped into 18 clusters according to their transcript abundance patterns across all samples, and plotted using a heatmap to visualise their expression patterns. Figure 5.5 illustrates the expression pattern of the groups of samples. Yellow, orange, red and brown colours indicate strongly up-regulated genes, green indicates that the clustered genes changed

very little, and turquoise and blue shades indicate that genes in the cluster were strongly down-regulated (Figure 5.5). For instance, when comparing tip and base samples (TvsB), genes that were abundant transcripts in the tip are in yellow, orange, red and brown colours, and the blue shades represent genes with higher transcript levels in the base samples (low expression in tip). From the results of the scoping experiments on transcript abundance in different leaf ages and leaf sections sampled over the light/ dark cycle and investigated for CAM, clock and sugar metabolism gene transcript levels using RT-PCR, and the metabolic and physiological analysis presented in chapters 3 and 4, it was clear that the leaf tip performed CAM, whereas the pale green leaf base used C_3 photosynthesis. With these preliminary and supporting correlated results as a guide, the subsequent analysis of the clusters of DE genes focussed on the genes with higher transcript levels in the leaf tip relative to the C_3 base and the non-photosynthetic white basal section of the leaf, as those genes were the strongest candidates for performing functions associated with optimal CAM in the *A. sisalana* leaf tip. The genes in clusters 10 - 18 were mostly highlighted in green and yellow shades with some red-brown shades. This indicated that these clusters of genes, which were highly expressed in the leaf tip, were candidates with potential roles in CAM. Thus, the genes in clusters 10 – 18 were selected for further investigation.

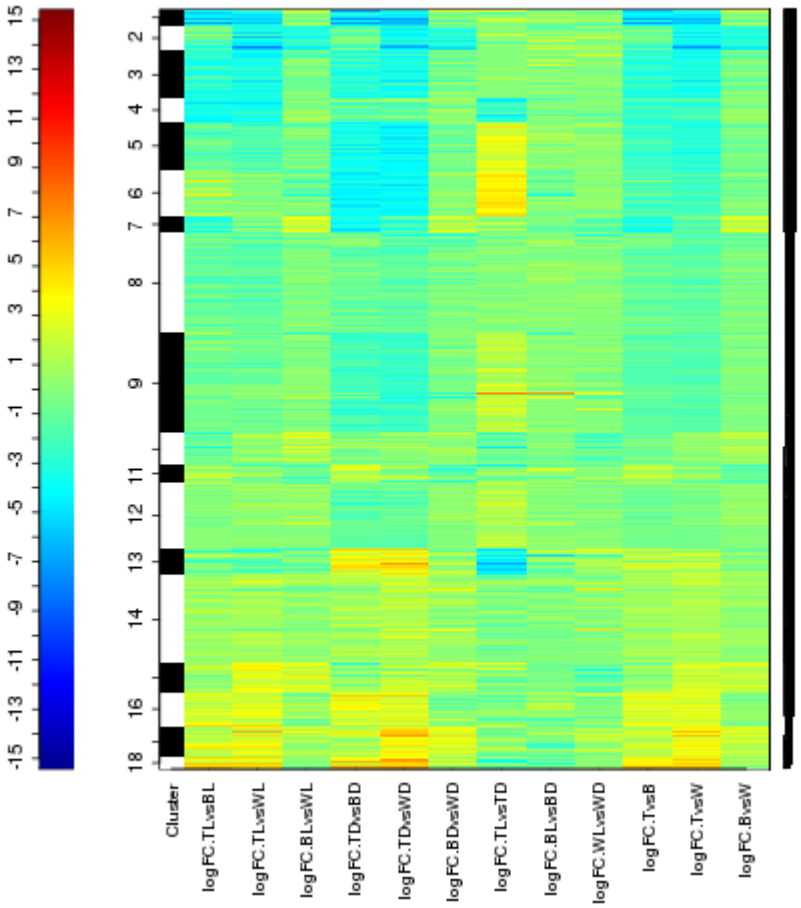


Figure 5.5 Differential gene expression between leaf segments and time points in 18 clusters.

Gene cluster analysis is presented as a heatmap. Expression levels were determined by Log-fold change (LogFC) values calculated from edgeR analysis. The different colours represent different LogFC values as indicated in the multi-coloured bar to the left with a range of values. The abbreviation of group of the youngest fully expanded *A. sisalana* leaf samples are named as follows; TL = Tip Light (10:00), TD = Tip dark 22:00), BL = Base Light, BD = Base Dark, WL = White Light, WD = White Dark. Each group represents 3 biological replicates of samples. T = Tip, B = base and W = White represent the total transcripts of light and dark samples in each leaf segment to allow the comparison regardless of sampling time. Versus (vs) indicates the pairwise comparison of groups of samples. The black vertical bar on the right indicates the DE genes, but due to the extremely high number of genes represented in this analysis, this bar does not provide any visible details. The heatmap was generated by Dr. Yongxiang Fang, CGR, University of Liverpool.

The DE genes were sorted using an Excel programme in order to identify genes of interest that fulfilled the selection criteria that identified them as candidate CAM-associated genes; it had been established in the previous chapters that the leaf tip performed full CAM, the pale green base performed C_3, and the white basal section of the leaf was most likely non-photosynthetic

leaf tissue. A total of 19,353 DE genes were identified as being significantly more abundant in the leaf tip compared to the pale green base and the basal white section with false discovery rate (FDR: corrected and adjusted P-values) of < 0.05. The sorting of these ~19,000 genes focused first on a comparison of genes that were most abundant in the leaf tip compared to the pale green base, as these genes were the ones most likely to differentiate between CAM in the tip and C_3 in the pale green base. Thus, genes showing higher transcript abundance in the tip relative to the base were retained, while those that were only higher in the tip compared to the white were excluded. The results of this filtering of the list of tip up-regulated genes revealed that most of the genes that were more abundant in the leaf tip relative to the base were also more abundant in the tip when compared to the white basal leaf tissue. Out of 19,353 DE genes, 2,457 genes only were higher in the leaf tip compared to the base, whilst 16,896 genes were more abundant in the leaf tip relative to both the pale green C_3 base and white basal section of the leaf (Figure 5.6B). These tip-abundant transcripts were ranked based on their Log-fold change (LogFC) sorting them from the most strongly tip-induced genes to the least tip enhanced genes. Most of the highly tip enhanced DE genes had a very small FDR value.

Genes in cluster 10-18 from Figure 5.5 were contrasted using Venn diagrams with the 19,353 tip-enhanced genes (Figure 5.6). Perhaps not surprisingly, the Venn diagrams showed that all of the tip-enhanced DE genes were represented by the genes in clusters 10 - 18 (Figure 5.6A), which further emphasised that these lists of genes were most likely to include the majority of the candidate CAM genes. The set of DE 19,353 tip-enhanced genes was also compared using a Venn diagram with the genes that had higher expression in the tip sampled in the light compared to the dark, and vice versa. The Venn diagram revealed that 3,278 genes were more highly expressed in the leaf tip sampled in the dark compared to the light (Figure 5.6C). Furthermore, 539 genes showed higher expression in the tip light samples compared to the

dark (Figure 5.6D). Interestingly, the number of genes with higher expression in tip dark was noticeably higher than the number of genes that were highly expressed in tip light. This indicated that considerably more of the tip-enhanced DE genes were most abundant in the dark. However, it is important to emphasise that a large proportion of tip-enhanced genes (15,536) did not show any difference in their transcript abundance in terms of light/ dark regulation.

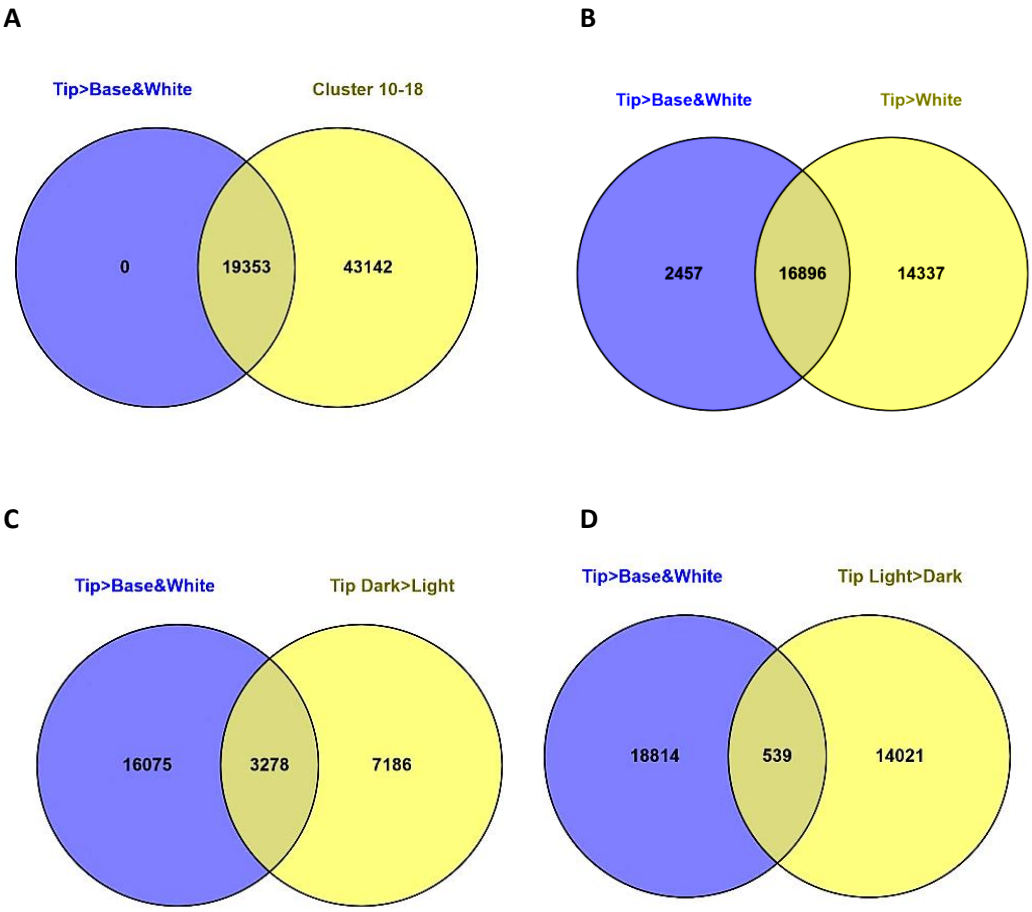


Figure 5.6 Differential gene expression between leaf segments and time points.

The numbers of genes differentially expressed or shared among different sets of DE genes. Genes highly expressed in leaf tip were compared using Venn diagrams with genes in clusters 10 - 18 from Figure 5.5 (A), all genes expressed higher in the leaf tip compared to white basal segment only (B), genes expressed higher in leaf tip sampled in the dark (22:00) compared to the light (10:00) (C), and genes expressed higher in the leaf tip sampled in the light compared to the dark (D). The diagrams were generated using online Venn diagram generator. FDR was < 0.05 for all DE genes compared using edgeR.

The 19,353 tip-enhanced DE genes were then continued into functional analysis using R to measure the frequency of annotations of genes in that set of genes, based on their annotation against the Pfam database and their Gene Ontology (GO) annotation. The 30 most frequent functionally annotated genes were obtained and plotted in Figure 5.7 and 5.8. It was also found that these top 30 tip-enhanced gene annotations were also ranked at the top of the list of genes that contained high logFC values. In terms of the GO annotation enrichment analysis, a number of genes were annotated with functions involved in regulatory processes including DNA-binding transcription factors, protein binding, and protein kinases (Figure 5.7). Other genes were associated with catalytic activity, membrane localised transporter and other functions (Figure 5.7). In *A. deserti,* GO-enrichment analysis showed that the leaf tip contained higher transcript levels of genes associated with photosynthesis, chlorophyll biosynthesis, and additional regulatory proteins (Gross *et al*., 2013). Furthermore, genes associated with processes associated with growth and cellular development were enhanced in the leaf basal segments, whilst several important energy-related metabolic processes were enhanced in the leaf tip (Gross *et al.*, 2013).

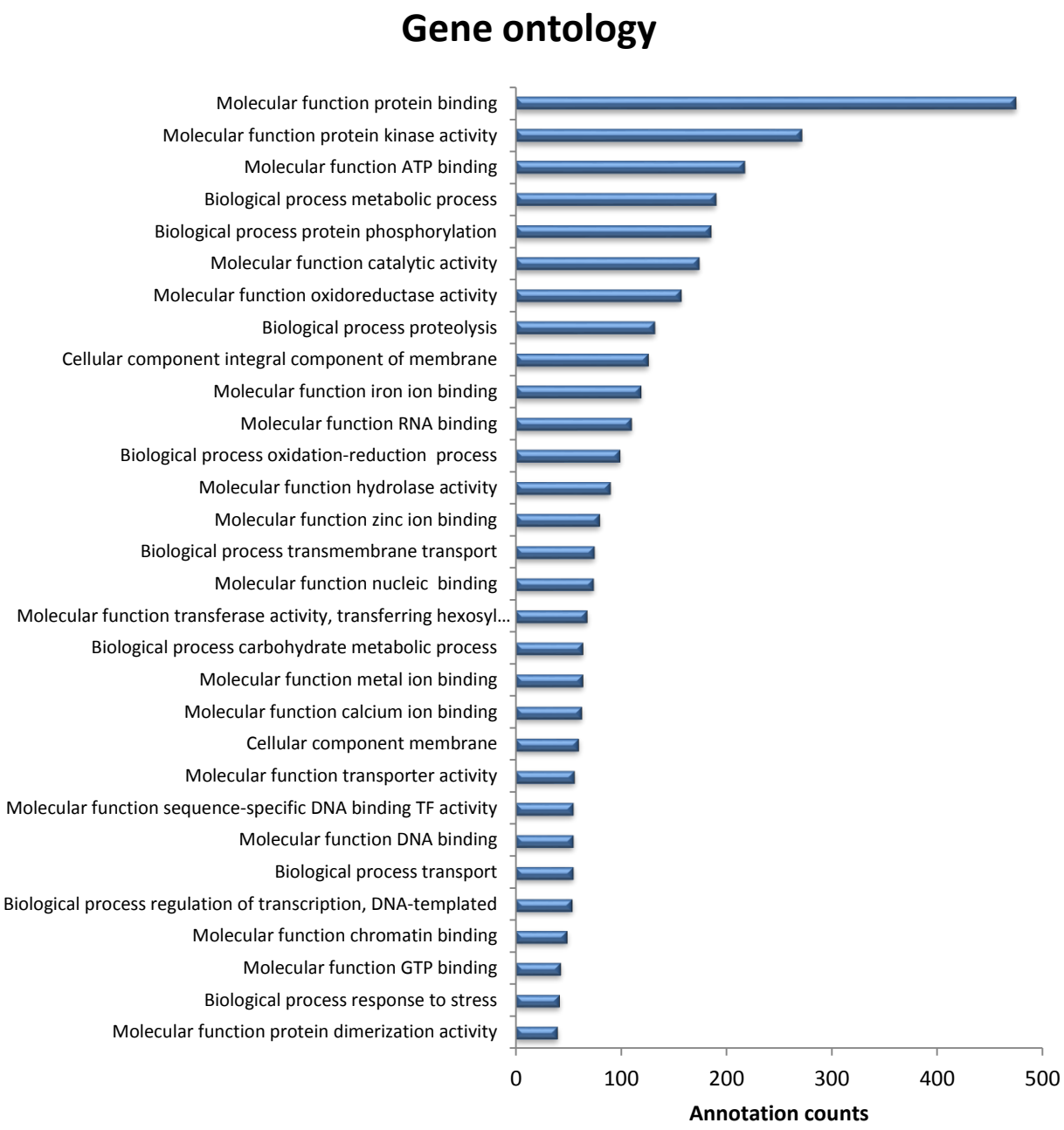


Figure 5.7 The top 30 most frequent functional annotations from all the genes that were differentially expressed in the leaf samples based on Gene Ontology (GO) database annotation.

The tip-enhanced DE genes were analysed using R to count the number of gene annotations that occurred in that set of DE genes based on their GO database annotation. The 30 most frequent functionally annotated groups of genes were obtained. Y-axis represents GO protein function. X-axis represents annotation counts.

Based on annotation of predicted gene function using the Pfam database (Figure 5.8), the analysis of the same DE set of genes ranked cytochrome P450 proteins (CYPs) as the most frequently annotated genes with 89 counts, followed by protein kinase domain and tyrosine kinase (Figure 5.8). These proteins are commonly found in plants. A group of sugar metabolism related genes, annotated as 'sugar transporter', was also a frequent gene annotation amongst the list of tip-enhanced genes, with 15 counts for this annotation category. Photosynthesis-related genes, including chlorophyll A-B binding proteins, and importantly the CAM-associated gene family, malate dehydrogenase (MDH), were also amongst the top 30 genes functional annotations in the list of tip-enhanced DE genes (Figure 5.8). Photosynthetic genes were not frequently and highly represented in these enriched annotation categories and groups, probably due to the fact that the DE genes analysed were abundant in the CAM-performing leaf tip compared to the C_3 pale green leaf base, with both tissues sharing a number of common photosynthetic genes (Figure 5.8). However, when comparing the photosynthetic leaf tissues with the non-photosynthetic tissues, namely the white basal leaf section, a number of photosynthetic gene annotations including 'photosynthetic reaction centre', and 'chlorophyll A-B binding protein family' were amongst the most frequent gene annotations in the leaf tip and base, while no photosynthesis-related gene annotations were found to be abundant amongst the list of genes that were abundant in the white basal leaf tissues (data not shown). Thus, these findings taken together with the CO_2 exchange IRGA results in Section 4.2.1 support the proposal that the leaf base of these youngest, fully-expanded *A. sisalana* leaves perform C_3 photosynthesis, whereas the white basal leaf section had yet to undergo photomorphogenesis and become photosynthetically active.

Interestingly, the list of genes that were highly differentially expressed in the leaf base also ranked Cytochromes P450 (CYPs) as the most frequent gene annotation; the same group of genes that appeared at the top of the tip-enhanced list. CYP is the largest protein family in

plants, and these proteins have been found to be involved in a wide range of biosynthetic reactions, including the catabolism of plant bioactive molecules (Morant *et al.*, 2003). Most of the oxidation steps in plant secondary metabolism are catalysed by CYP proteins leading not only to simple hydroxylations or epoxidations, but also to more diverse and complex reactions (Kahn and Durst, 2000; Werck-Reichhart *et al.*, 2002) which are required for the production of the valuable plant natural products with nutritional value (Morant *et al.*, 2003). The second from the top ranking of base-enhanced genes was cellulose synthase and MYB-like DNA-binding domain proteins which are transcription factors involved in gene regulation. Apart from cellulose synthase, several sugar/ carbohydrate-metabolism-related genes such as 'plant invertase methylesterase inhibitor' (PMEI), 'glycoside hydrolase family 176', and 'glycosyl transferase family 8' were also found amongst the base-enhanced genes. The presence of these abundant genes in the leaf base is supported by the idea that Agave leaf base is a sink tissue, where long-term-storage carbohydrates are synthesised and stored.

Protein kinases are known to be one of the most abundant protein families in higher plants and they have a very wide range of regulatory functions within the cell (Lehti-Shiu and Shiu, 2012), and tyrosine kinases, a subset of protein kinase family, are also known to be abundant in plant genomes (de la Fuente van Bentem and Hirt, 2009). However, in this study, these two protein kinase categories were not found to be frequently annotated amongst the DE genes in the pale green C_3 leaf base (data not shown), whereas these gene annotations were ranked amongst the top 3 most frequent gene annotations for the DE genes enhanced in abundance in the leaf tip and the white basal leaf section. This suggests that protein kinases were not as strongly differentially regulated in the pale green leaf base compared to the other parts of the leaf. As protein kinases function in the post-translational regulation of other enzymes that could have diverse functions, further more detailed analysis is needed to study the functions of the enzymes phosphorylated by these base-down-regulated protein kinases. In the white

basal leaf section, CYPs were not ranked as a frequent gene annotation amongst the DE genes, whereas CYPs were common amongst the list of DE genes identified as being enhanced in the leaf tip and the pale green leaf base. Genes associated with the carbohydrate metabolism related genes such as 'glycoside hydrolase family 17' and 'X8 domain' were ranked amongst the top highly expressed genes in white basal section of the leaf (data not shown). This might possibly indicate that white basal section of the leaf can be also sink tissues where long-term-storage carbohydrates are synthesised and stored.

Some of the transcription factor (TF) families in the Pfam list (Figure 5.8) were also found in the transcriptome study of two other, related *Agave* species, namely *A. tequilana* and *A. deserti* (Gross *et al*., 2013). In *A. deserti*, GRAS, YABBY, MYB, bHLH, and Zn finger transcription factor families were identified as being highly differentially expressed in the leaf base, and the KNOX class I TFs were reported to have strongest expression in the leaf base (Gross *et al.*, 2013). In this study, members of the MYB TF family were highlighted as being highly differentially expressed in all leaf tissues of *A. sisalana* (base and white, data not shown; for tip, see Figure 5.8). The YABBY and KNOX class I annotations were, however, not detected as being abundant gene annotation categories in leaf base-enhanced DE genes (data not shown). Instead of being detected as DE genes in the pale green leaf base, a small number of KNOX class I family TF genes were detected as DE genes in the white basal segment of the *A. sisalana* leaf (data not show). This difference may be due to the fact that the leaf base in this *A. sisalana* study was sampled from pale green basal tissues, whilst the white basal section of the leaf was sampled as solely white tissue from the very base of the leaf (numbered 1 in Figure 2.2). In contrast, the *A. deserti* leaf base samples used in the study by Gross *et al.*, (2013) included the whole leaf base, which consisted of the white part plus the pale green part as a single sample. Hence, the white segment reported here is comparable to leaf base in the study of Gross *et al.*, (2013), both having KNOX class I family TF genes amongst the list of DE genes

that were detected to have enhanced transcript abundance in the leaf base in comparison to the leaf tip.

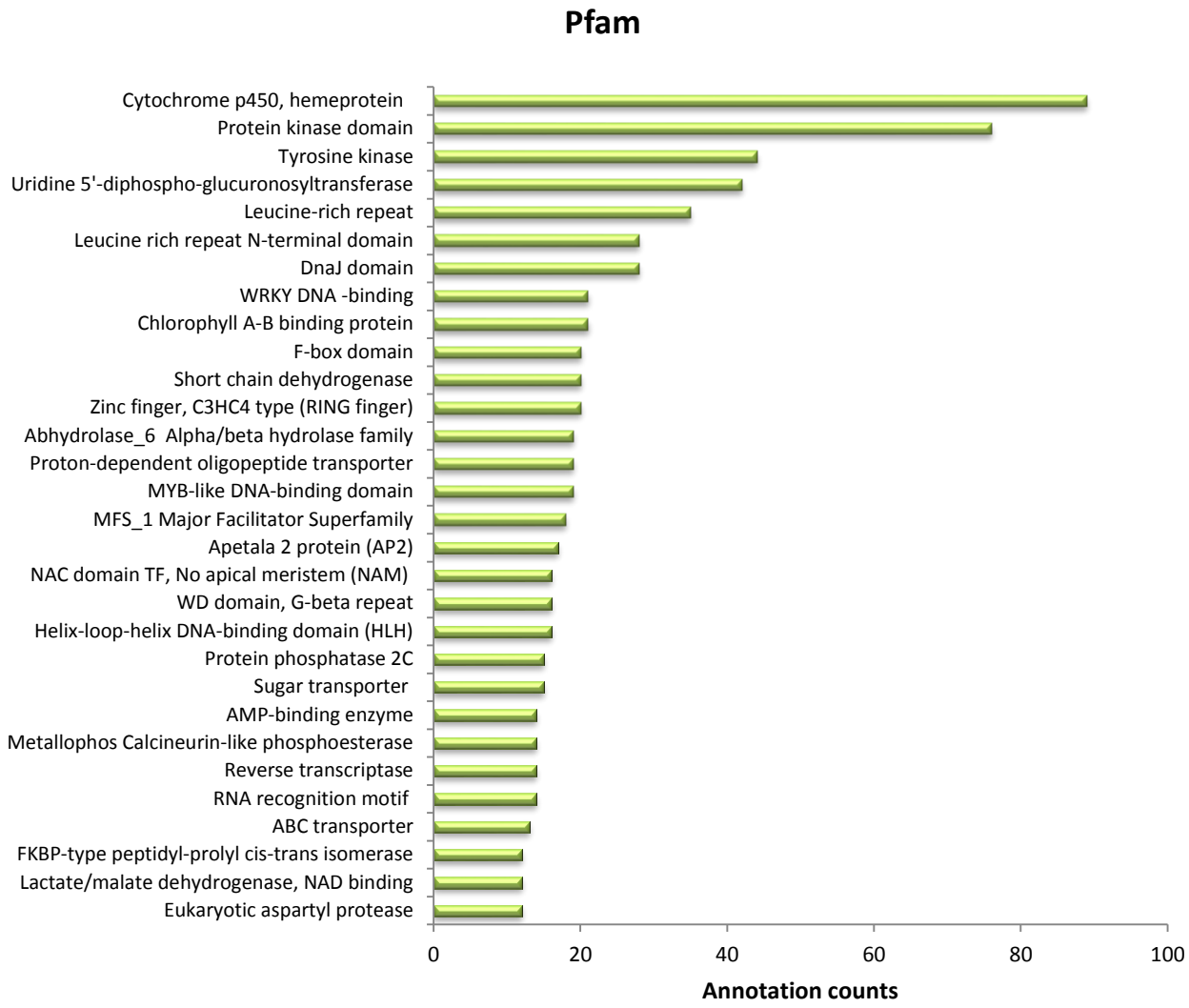


Figure 5.8 The top 30 most frequent function annotations of DE genes that were highly expressed in the leaf tip samples, based on Pfam database annotation of the DE genes.

The tip-enhanced DE genes were analysed using R to count the number of gene annotations, based on Pfam database annotation, of each category that occurred in the set of DE genes that showed enhanced transcript levels in the leaf tip. The 30 most frequent functional annotations are displayed. Y-axis represents Pfam protein names. X-axis represents annotation counts.

KNOX class II TFs, which tend to have wide patterns of expression, and MADS-box transcription factors, which regulate various developmental processes, were detected as being more abundant transcripts in the leaf tip relative to the leaf base in *A. deserti* (Gross *et al.*, 2013). By contrast, KNOX class II genes were only rarely detected, and MADS-box genes were not found amongst the most strongly DE genes in the *A. sisalana* leaf tip in this study. However, Li *et al.*, (2010), found MADS-box genes were highly expressed in the leaf base in monocot grass maize (*Zea mays*). These contrasting patterns of MADS-box gene regulation in different studies might be due to the fact that MADS-box genes can differ broadly in expression and function in the Agave leaf, but their role in metabolic processes and leaf developmental biology is still unknown (Delgado Sandoval Sdel *et al.*, 2012). MADS-box genes may also have different functions in different species, and different members of the MADS-box family will perform different gene regulatory, cell signalling and developmental control functions throughout the plant. Homeobox TF family genes were also highly abundant in the leaf base in maize (*Zea mays*), where they were proposed to perform diverse functions in leaf development (Li *et al.*, 2010). This is consistent with the results obtained in this study, where Homeobox genes were found to be most abundant in the *A. sisalana* leaf base (data not shown). GRAS, bHLH, and Zn finger TFs were low abundance transcripts in the leaf base (data not shown) whilst bHLH and Zn finger TFs were highly expressed in leaf tip in this study (Figure 5.8). Individual transcription factors from the same family can be expressed in different parts of the leaf. Li *et al.*, (2010) reported for maize leaf development that several transcription factors which belonged to the same family, were highly expressed in different leaf segments. NAC and MYB, for example, were reported to have various family members highly expressed across the whole leaf mainly in leaf base and tip with a few expressed in the transitional area between base and tip. Golden 2 (G2) and DNA-binding with one finger (DOF) TFs have been shown to regulate photosynthetic gene expression and were found to be tip enhanced TFs in the maize study (Li

et al., 2010). Several NAC and DOF family TFs were also found to be enhanced in transcript abundance in the leaf tip in this study (Table 5.3).

Selection of candidate CAM-associated transcription factors

In the dicot CAM species, *K. fedtschenkoi*, a number of core CAM pathway genes are among the most strongly differentially regulated genes found in full CAM performing leaves when compared to C_3 leaves (Boxall, Dever and Hartwell, personal communication). Gross *et al.*, (2013) also reported that known CAM-associated genes, including *PPC*, *MDH* and *PPDK*, were detected as induced transcripts in the *A. deserti* leaf tip, along with other highly expressed photosynthetic genes. In this study, several known CAM pathway genes were also found among the top genes that were strongly expressed in *A. sisalana* leaf tip in terms of expression level (logFC) (Supplementary Table 5.1). Thus, the identification of novel CAM genes was focused on the top genes that were induced in the leaf tip compared to the pale green base and white basal sections.

Amongst the list of the most strongly tip-enhanced genes, regulatory genes such as transcription factors were discovered through a simple word search within their annotation information using a Linux script and Excel. The FPKM transcript abundance values for each identified TF gene were used to visualise the expression pattern across samples (Figure 5.9 and Figure 5.10). TFs were selected due to the fact that they are one of the most important known categories of regulatory genes that play an important role in the regulation of downstream genes, controlling the level of transcription of their target genes (Latchman, 1997; Karin, 1990). Certain contigs annotated as encoding novel TF genes, having similar regulatory patterns to known CAM genes, and FPKM values equal to or higher than 10, were selected as candidate TF genes (Table 5.3). In addition to the leaf development, the light/ dark pattern of regulation of each identified TF was also considered when selecting candidate genes.

Table 5.3 Candidate genes encoding novel transcription factors (TF) selected from the list of the most strongly tip-enhanced DE genes (CAM potential) and a control non-CAM gene highly expressed in white segment of leaf.

Acronym : Contig name	Gene names/description
CAM potential	
AsNAC : c566713_g1	NAC domain, No apical meristem (NAM) TF
AsWRKY : c571790_g2	WRKY DNA -binding protein TF
AsPLATZ : c541787_g1	Plant AT-rich sequence and zinc-binding proteins TF
AsBTB : c599899_g1	BTB/POZ protein TF
AsAP2 : c582092_g1	Apetala 2 protein TF
As_zf_DOF : c534926_g1	DOF protein domain, zinc finger TF
AsHomeobox : c526089_g4	Homeobox protein TF
Non-CAM	
Class I *AsKNOX1* : c568644_g2	Knotted1-like homeobox protein TF

Among the novel candidate genes, *AsNAC* (c566713_g1)*, AsWRKY* (c571790_*g2*), and *AsAP2* (c582092_g1) were also found in the list of the top 30 tip-enhanced most frequent functionally annotated genes (Figure 5.8). Most of the other candidate genes were also frequently found in the tip-expressed list, apart from the top 30 most frequent function annotations (data not shown). In addition, a non-photosynthetic gene, class I *AsKNOX1* (c568644_g2), was also selected from the list of 9,048 genes that were most strongly expressed in the basal white section of the leaf when compared to both the tip and pale green base leaf section. Known CAM genes also were used as a control for CAM induction and plotted using FPKM values to observe the regulation of their transcript abundance between the different leaf sections and the light and dark samples (Figure 5.10). Individual bar charts were plotted for each gene using the mean FPKM values (Figure 5.9 and 5.10). These charts show that *AsNAC* (c566713_g1) and *AsWRKY* (c571790_*g2*) exhibited remarkably strong induction of their transcript level in the leaf tip sampled in the dark, with very low transcript levels in the other samples (Figure 5.13A

and B). This correlated well with the fact that these genes appeared amongst the list of the most strongly tip-enhanced and frequently annotated genes (Figure 5.8).

AsPLATZ (c541787_g1) showed very strong induction in the leaf tip relative to the other leaf sections (Figure 5.9C). However, *PLATZ* family TFs were not as abundant as other transcription factors in the annotation enrichment analysis (data not shown), although this may be because the *PLATZ* gene family is not as large as many of the other TF gene families and thus there may be fewer genes in the *A. sisalana* genome that could be identified as strongly DE in the leaf tip relative to the base and white tissue. *AsBTB* (c599899_g1) showed higher transcript levels in the tip light than the tip dark (Figure 5.9D). In addition, *AsAP2* (c582092_g1) and *As_zf_DOF* (c534926_g1) were also selected for further more detailed analysis based on their distinctive patterns of regulation across the leaf developmental gradient (Figure 5.9E and F). *AsAP2* (c582092_g1) was detected at very similar transcript levels in the light for all three leaf segments, whereas in the dark it was virtually undetectable in the pale green base and white tissue, but reached its peak abundance in the leaf tip, being almost 4-fold higher in the dark in the tip than in the light in the tip (Figure 5.9E). *As_zf_DOF (c534926_g1)* was highest in the tip both in the light and the dark, but was much higher in the tip in the dark than the tip in the light (Figure 5.9F). Furthermore, *As_zf_DOF (c534926_g1)* showed a steady decline from the tip to the pale green base to the white basal tissue, although it was always higher in the dark than the light for all three leaf sections (Figure 5.9F).

AsHomeobox (c526089_g4) was also chosen as it showed high transcript levels in the leaf tip in the light, and displayed a strong light/ dark pattern of regulation in the leaf tip (higher in the light) which was reversed in the white basal part of the leaf such that its transcript abundance was higher at 22:00 dark in the white leaf base (Figure 5.10A). A number of Homeobox families TF genes were also detected as being more abundant in the leaf base in *A. sisalana* (data not

shown). The transcript levels of these base-enhanced *AsHomebox* genes were higher in the light than dark samples (data not shown).

Class I *AsKNOX1* (c568644_g2) was selected as a white basal section enhanced positive control gene (Figure 5.10B). When aligned with Arabidopsis database (Lamesch *et al.*, (2011); https://www.arabi-dopsis.org/Blast/), class I *AsKNOX1* (c568644_g2) was most closely related to Arabidopsis class I KNOX protein (AT1G62360.1). This protein is involved in shoot apical meristem formation during embryogenesis and also functions of apical meristem throughout the lifetime of the plant. This is very well consistent with the findings in this study for the regulation of the class I *AsKNOX1* (c568644_g2) in white basal section, which stages of leaf development take place, including cell division while it was low to undetectable in the older sections of the leaf that had undergone photomorphogenesis (Figure 5.10B). This class I *AsKNOX1* (c568644_g2) result for *A. sisalana* is also consistent with the findings of Zhou *et al.,* (2012), who demonstrated that a class I *knotted*-like homeobox gene (*Asknox*) was an abundant transcript (both in terms of Sanger EST sequencing and real-time PCR) in the apical meristem, which is located below the white basal part of the leaf. The *Asknox* gene was present at a much lower transcript level in the other leaf tissues, and undetectable in mature leaves of *A. sisalana* (Zhou *et al*., 2012). This finding is also similar to the transcript abundance profile of *KNOX* family genes in other species (Hake *et al.*, 2004; Hay and Tsiantis, 2009).

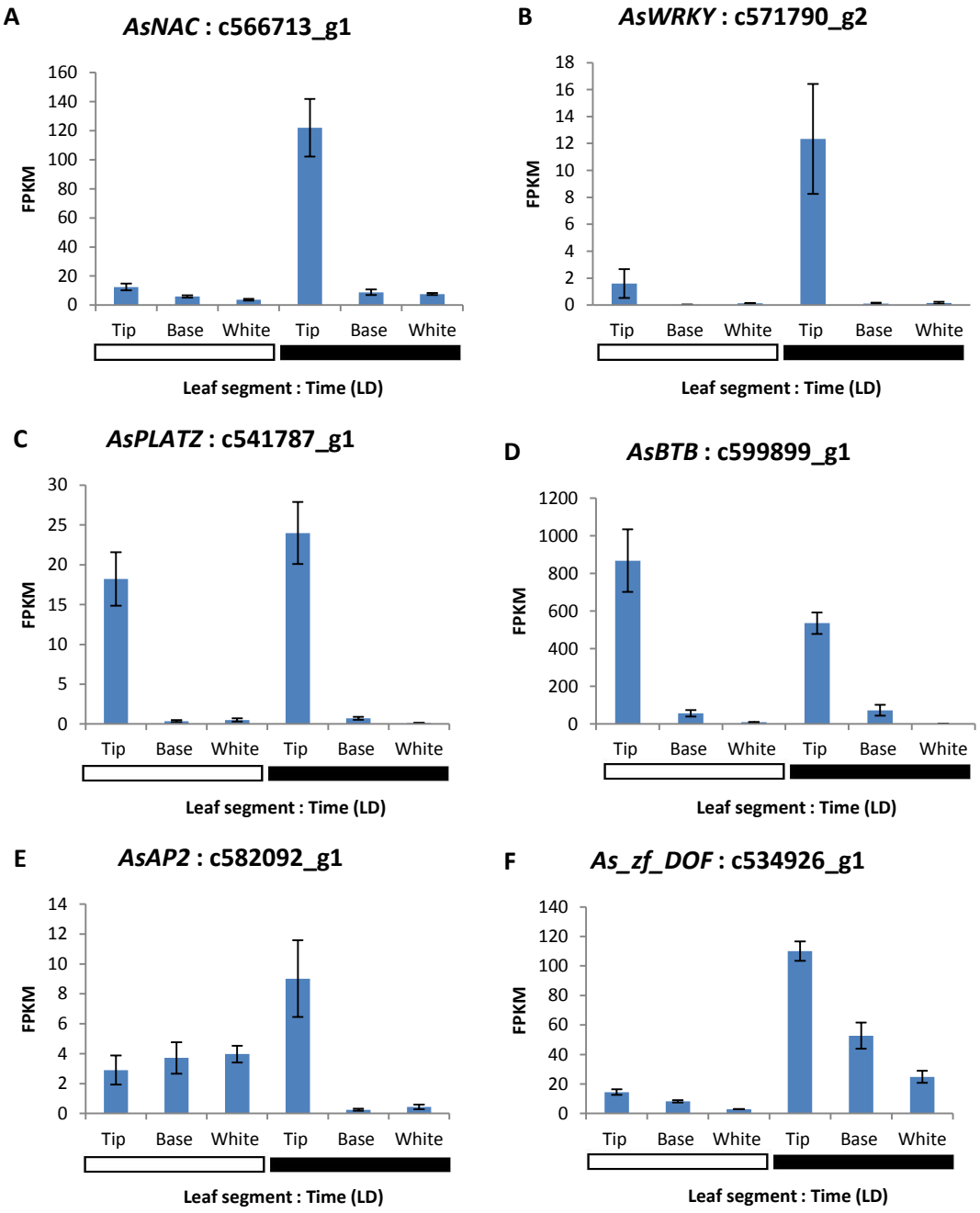


Figure 5.9 Identification of novel, candidate CAM-associated transcription factors that are induced in the leaf tip section of *A. sisalana* leaves.

The mean FPKM values from the Illumina RNA-seq dataset were plotted for several novel candidate CAM-associated transcription factor genes. The graphs present the transcript abundance values (FPKM) for: *AsNAC* (A), *AsWRKY* (B), *AsPLATZ* (C), *AsBTB* (D), *AsAP2* (E), *As_zf_DOF* (F). The white bar indicates light (10:00, 2 h before dusk), and the black bar indicates dark (22:00, 2 h before dawn).

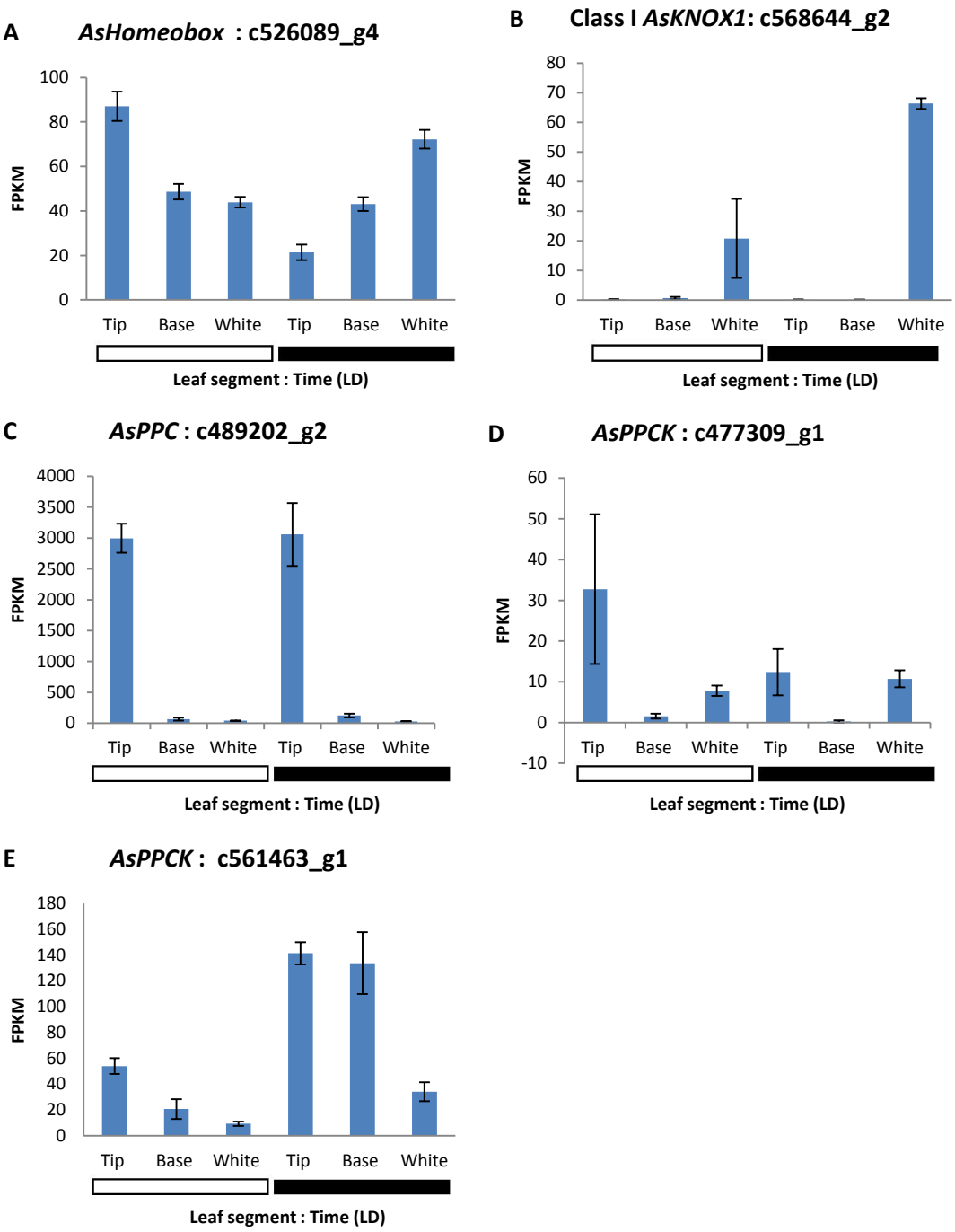


Figure 5.10 Identification of novel, candidate CAM-associated, non-photosynthetic and known CAM transcription factors that are induced in the leaf tip section of *A. sisalana* leaves.

The mean FPKM values from the Illumina RNA-seq dataset were plotted and the graphs present the transcript abundance values (FPKM) for: novel candidate CAM-associated *AsHomeobox* (A), non-photosynthetic Class I *AsKNOX* (B), and known CAM transcription factor genes *AsPPC* (C) and *AsPPCK* (D & E). The white bar indicates light (10:00, 2 h before dusk), and the black bar indicates dark (22:00, 2 h before dawn).

In addition to the selection of the novel candidate CAM-associated TF genes, 3 control CAM genes were also plotted for direct comparison to the regulation of the selected TFs (Figure 5.10C, D and E). As expected, *AsPPC* (c489202_g2), which encodes the primary CAM carboxylase that catalyses dark period CO_2 fixation leading to malate accumulation in the vacuole, displayed a very high transcript abundance in the leaf tip relative to the other leaf sections (Figure 5.10C). This pattern of *AsPPC* transcript abundance corresponded well with the other previous semi-quantiative RT-PCR results presented in Chapter 3, in which the primers used in the semi-quantitative RT-PCR targeted the same *AsPPC* contig (c489202_g2) presented in this chapter. The transcript abundance of *AsPPC* in the CAM performing leaf tip varied very little between the light and the dark sample (Figure 5.10C). The *AsPPCK* (c477309_g1 and c561463_g1) genes also exhibited greater transcript abundance in the leaf tip relative to the pale green base, both in the light and the dark with the difference between leaf tip and base being more pronounced for contig *c477309_g1* than for contig *c561463_g1* (Figure 5.10D and E). However, the difference was less marked when comparing the leaf tip with the basal white samples, particularly during the dark period especially for contig *c561463*_g1 (Figure 5.10E). In terms of light/dark regulation of *AsPPCK*, the transcript abundance of contig *c477309_g1* (Figure 5.10D) was higher in the tip-light sample than the tip-dark sample, which contrasts with the regulation of contig *c561463_g1* (Figure 5.10E), and the well-known pattern of *PPCK* regulation in other CAM species whereby *PPCK* transcript levels rise during the dark period under the control of the central circadian clock (Hartwell *et al*., 1999; Boxall *et al*., 2005; Dever *et al*., 2015). This pattern for the regulation of *AsPPCK* (c477309_g1) in terms of leaf development and light/dark regulation according to the RNA-seq dataset does not correlate particularly well with the phosphorylation state of PEPC according to the immune-blot analysis in chapter 4, which demonstrated that PEPC was only phosphorylated in the tip dark samples with no phosphorylated PEPC detected in any of the

other leaf sections either in the light or the dark (Figure 4.3 and 4.4). However, as post-transcriptional and translational regulation also play a major role of gene expression, the transcript level of *AsPPCK* (c477309_g1) found here might not necessarily reflect the protein level and activity of *PPCK*. In addition, it should be noted that the error bar for *AsPPCK* (c477309_g1) transcript abundance for the tip light sample was large signifying that variation in the transcript level between the 3 biological replicates was large, which might make the comparison between tip light and dark more error prone (Figure 5.10D). More than one *PPCKs* have been reported in C_3 rice (Fukayama *et al.*, 2006) and C_4 monocot crop maize (Zea mays) (Shenton *et al.*, 2006); these novel *PPCK*s showed different expression patterns, the leaf cells and clear different roles in the regulation of PEPC phosphorylation state. Thus, there may be other *AsPPCKs* with different transcript abundance profiles in *A. sisalana*, for example, PPCK may also be functional in the white basal tissues (Figure 5.10D). Another possibility is that the pattern of *AsPPCK* transcript abundance regulation in *A. sisalana* may be different from the previously defined patterns of *PPCK* regulation in other CAM species, even though the phosphorylation of PEPC in the *A. sisalana* CAM leaf tip clearly performs the classic pattern of dark period phosphorylation. However, it might be too early to draw conclusion regarding the precise pattern of *AsPPCK* regulation in *A. sisalana* at this stage, especially considering the fact that only two time points were sampled representing the dark and the light period. *AsPPCK* transcript levels tend to peak in the middle of the dark period in other CAM species, and so it may be that the dark sample used here for the RNA-seq, which was sampled 2 h before dawn, failed to capture the main peak of *AsPPCK* transcript abundance, as it may be that *AsPPCK* transcript peak much earlier in the dark period in *A. sisalana* and then decline such that they are low and/ or highly variable by 2 h before dawn. Further experiments are required to resolve these questions and possibilities. The results from the Q-RT-PCR analysis of candidate

CAM gene transcript abundance measured over the full 24 h LD time course experiment with samples at 4 h intervals for all candidate and control genes are described in Chapter 6.

5.3 Summary

This chapter has presented and discussed the results of a large scale RNA-seq experiment on different sections of the youngest, fully-expanded *A. sisalana* leaf sampled both in the light and the dark. Overall, the study produced a biologically meaningful dataset that defines the transcriptome-wide regulatory patterns for the different leaf sections both in the light and the dark. The sequencing worked well and generated a large transcriptomic data for the subsequent downstream analyses. *De novo* assembly was performed using the Trinity pipeline (Grabherr *et al.*, 2011), a tool recommended by the CGR, and generated a good quality transcriptome assembly, although the average contig length was not as high as for the assembled transcripts reported previously for *A. tequilana* and *A. deserti* (Gross *et al*., 2013). Trinity is equipped with an algorithm that allows it to detect paralogous genes and alternatively spliced transcript isoforms. Consequently, it assembled a large number of 671,886 'Trinity genes', which were even more numerous (941,989 isoforms) when alternatively spliced transcripts were taken into account (Table 5.1). The number of contigs recovered here, including the 118,807 protein-coding sequences found in this study, was a lot higher than the previously published *A. tequilana* and *A. deserti* transcriptomes (Gross *et al*., 2013). When the *A. sisalana* Trinity assembly was evaluated using CEGMA to investigate the conservation of the core eukaryotic gene set, 99.13% of the most highly conserved core eukaryotic genes were found to be represented in the assembled transcriptome. Moreover, 95.16% and 99.19% of ultra-conserved core eukaryotic genes were completely and partially present in the transcriptome respectively. Hence, these parameters suggest that the transcriptome assembly reported here represents a high quality transcriptome for *A. sisalana*

as it included a high percentage conserved genes, which suggests a high degree of completeness of the transcriptome coverage.

Differential expression analysis was performed at the level of the 'Trinity genes' using the 671,886 assembled genes. Each gene consisted of one or more alternatively spliced transcript isoforms that shared sequence content. The 'Trinity genes' were used as the references sequence for the Bowtie2 mapping of the raw Illumina reads from all samples, and the mapped reads were counted per gene. The count data was used for the calculation of the differentially expressed gene set using edgeR (Robinson *et al.*, 2010). This initial stage of the downstream analysis revealed clear variation between the different leaf samples including the leaf tip, base and white sections. Principal component analysis and a Pearson correlation coefficient plot visualised via a heatmap showed that the difference in transcript abundance between the CAM performing leaf tip, especially in dark samples, and both C_3 leaf base and non-photosynthetic white leaf segment was a lot higher than the difference in transcript levels between the base and white samples (Figure 5.3, 5.4 and 5.5).

The DE genes were sorted and clustered based on their expression patterns. CAM genes were found among those transcripts that were highly abundant in the leaf tip compared to other segments. These DE genes had potential roles in the developmental induction and light/ dark and circadian clock mediated regulation of CAM, and were consequently identified and characterised in greater detail.

The functional annotation enrichment analysis ranked the 30 most frequent gene annotation categories of the most strongly differentially expressed genes in the leaf tip samples. This analysis identified novel genes and their functions that were highly expressed in different leaf segments. For the tip-abundant genes, GO annotation enrichment analysis revealed a number of genes annotated with functions associated with regulatory and cell signalling processes,

including transcription factors, protein binding, and protein kinases. Other enriched gene annotations included 'catalytic activity', 'transportation' and several other functions (Figure 5.7). Pfam database annotation enrichment analysis for the DE tip genes highlighted that 'CYP' was the most frequently annotated gene in the list, followed by 'protein kinase' and 'tyrosine kinase'. Several photosynthesis-related gene families were also amongst the top 30 Pfam category enriched annotation groups (Figure 5.8). CYP was also ranked at the top of the genes highly expressed in the pale green leaf base which performs C_3 photosynthesis, followed by 'cellulose synthase' and 'MYB-like DNA-binding domain'. Several sugar and carbohydrate related genes were also abundant categories in the DE tip gene list. Interestingly, protein inhibitor families were also found among the most strongly DE genes. However, the two protein kinase categories that were ranked amongst the top three most frequent annotation categories in tip and white basal leaf tissue DE genes were enriched categories in the DE gene list for the pale green leaf base samples. This indicated that 'protein kinases' were probably not as important to the biology of the pale green leaf base as fewer protein kinases were strongly differentially expressed in the leaf base when comparing to either the leaf tip or the white basal leaf tissue. In the white basal leaf tissue, CYPs were not ranked as an abundant category amongst the list of DE genes. Instead, in the white basal samples, genes associated with cellular functions and carbohydrate metabolism were enriched in the annotation of the DE genes. No families of photosynthetic genes were found to be enriched in the DE genes identified in the white basal leaf tissue, whereas photosynthetic genes were amongst the DE genes in the pale green leaf base and the dark green leaf tip.

In *A. deserti,* GRAS, YABBY, MYB, bHLH, KNOX class I, and Zinc finger transcription factor families were identified as being abundant transcripts in the leaf base (Gross *et al*., 2013). In this study, various groups of MYB family TFs were detected as highly expressed in all leaf tissues of *A. sisalana* (base and white data not shown; for tip, see Figure 5.8). However, the

YABBY and KNOX familes of TFs were not detected as being enriched functional annotation categories in the DE gene list for the leaf base (data not shown). A small number of class I KNOX family genes were differentially expressed in the white basal leaf tissue, possibly because the white basal section reported here is comparable to the leaf base used in the study of Gross *et al.,* (2013). KNOX class II and MADS-box TFs were highly abundant transcripts in the leaf tip in *A. deserti* (Gross *et al*., 2013), while few KNOX class II genes were in the DE gene list for the tip in this study, and MADS-box TFs were not found amongst the list of DE genes in the *A. sisalana* leaf tip. GRAS, bHLH, and Zinc finger were low abundance transcripts in the leaf base in this study (data not shown). In contrast, bHLH and Zinc finger TFs were highly expressed in leaf tip (Figure 5.8). Li *et al.*, (2010) found that Golden 2 (G2) and DNA-binding with one finger (DOF) transcription factors were up-regulated in the *Zea mays* leaf tip. Some zf_DOF family members were also found to be up-regulated in the *A. sisalana* leaf tip in this study (e.g. Figure 5.9F).

In this study, several CAM genes were also found among the top genes that were strongly up-regulated in the *A. sisalana* leaf tip relative to the pale green base and the white basal tissue (e.g. *AsPPC* and *AsPPCK*, Figure 5.10C, D and E). Seven novel transcription factors were identified that have the potential to play a role in the developmental induction of the CAM in the leaf tip, and/ or the light/ dark and circadian clock control of CAM in the tip. In addition, a non-CAM control gene was identified that was up-regulated in the white basal part of the leaf. FPKM values for each gene were used to illustrate graphically the transcript abundance levels of each gene in the different samples (Figure 5.10). Among the candidate genes selected, *AsNAC*, *AsWRKY* and *AsAP2* were also found in the list of the top 30 most frequently annotated genes within the DE gene list for the leaf tip. Transcript level based on FPKM values indicated that *AsNAC* and *AsWRKY* were up-regulated in the tip in the dark with very low levels in other samples (Figure 5.9A and B). *AsBTB* was more abundant in the tip sampled in the light than in

the dark (Figure 5.9D). The other genes were generally more abundant in the tip, but displayed a diverse range of patterns of regulations across the 6 samples. These transcript abundance patterns define these TF genes as interesting candidates that may function in the developmental induction and light/ dark and circadian clock control of CAM in the leaf tip in *A. sisalana*.

The class I *AsKNOX1*, a non-CAM control gene, was characterised by being abundant in the leaf white section but low-to-undetectable in other parts of the leaf (Figure 5.10B). This was consistent with the previous work of Zhou *et al.*, (2012), who also identified a class I *AsKNOX* gene that was expressed in the meristematic region of *A. sisalana*.

Very few studies have been published to date on Agave transcriptome analysis, and there is no whole genome sequence available for any *Agave* species. The results in this *A. sisalana* study have been compared with the results of Gross *et al.,* (2013); the only published Agave transcriptome study currently available. This study provides a comprehensive analysis of the regulation of the *A. sisalana* leaf transcriptome, and adds a valuable new whole trasncriptome analysis for a CAM species to the growing number of similar studies underway throughout the wider CAM research community. This study will also prove to be invaluable for the further understanding of the biology of Agaves, not only in terms of understanding CAM in Agave, but also for helping to reveal the functional genomics of the wide range of other biological adaptations that are found in Agaves. Perhaps most importantly, this work has provided a transcriptome-wide insight into CAM-associated genes and their light/ dark regulation, which will facilitate the further investigation of the biochemistry and functional genomics of CAM in this particular CAM species and other related species.

Due to the fact that the transcript levels presented in this chapter were only from RNA-seq analysis of samples collected at a single time point in the light and dark, analysis of the

regulation of some of the identified candidate CAM genes over the full 24 h LD time course was an important next step in the research. Q-RT-PCR analysis was performed using the 4 h interval light/ dark time course RNA samples from the white basal leaf tissue, pale green C_3 base section, and dark green CAM leaf tip, in order to further investigate the regulation of the novel candidate CAM genes throughout this expanded time course, and also to validate the RNA-seq results. The result of these Q-RT-PCR assays are described and discussed in Chapter 6. In addition, the 82 h LL circadian, free-running time course experiment with leaf tip samples collected every 4 h from fully-developed-CAM leaves was also investigated for the identified candidates CAM genes using Q-RT-PCR in order to determine whether or not the transcript abundance of the identified candidate CAM genes was under the control of the central circadian clock.

Chapter 6

A detailed Q-RT-PCR investigation into the light/ dark and circadian clock control of CAM-induced transcription factors identified using RNA-seq

6.1 Introduction

As described in chapter 3, early in this PhD project, semi-quantitative RT-PCR analysis was carried out for a series of scoping experiments that helped to determine the optimum sampling regime for the RNA-seq work described in chapter 5. Semi-quantitative RT-PCR was used for those experiments because of its relative affordability, and because of its ability to distinguish very readily the large changes in transcript abundance that occur for CAM genes both between the leaf base and the leaf tip and over the light/ dark cycle. It was also ideally suited to the large number of samples used in the light/ dark time course experiments. However, in order to achieve more detailed and accurate quantification of the transcript abundances of the novel transcription factor genes identified through the RNA-seq work in chapter 5, the quantitative real-time reverse transcriptase PCR (Q-RT-PCR) technique was applied to obtain the results presented in this chapter. Q-RT-PCR is readily automated and a straightforward technique to adapt for high throughput studies. It also achieves superior sensitivity and specificity of quantification when compared to the more traditional semi-quantitative RT-PCR that relies on end-point PCR that is quantified through the measurement of DNA band intensities on ethidium bromide stained agarose gels (Bustin, 2010). Q-RT-PCR has been used extensively in plant research including for the study of gene transcript

abundance profiles, and has increasingly become the method of choice for the accurate quantification of gene transcript levels, replacing other more conventional or traditional techniques such as Northern blotting and gel-based semi-quantitative RT-PCR (Gachon *et al.*, 2004).

The purpose of the Q-RT-PCR experiments described in this study was to study in greater detail the transcript abundance profiles of the CAM-induced transcription factors that were newly discovered via the *de novo* RNA-seq analysis described in Chapter 5. Specifically, the experiments in this chapter were aimed to extend understanding of the light/ dark regulation of each transcript beyond the single light (10:00, 2 h before dusk) and dark (22:00, 2 h before dawn) samples used for the RNA-seq experiment. Although a detailed time course of leaf developmental RNA samples, with samples collected every 4 h throughout a 24 h 12:12 light/ dark cycle, was collected prior to the RNA-seq experiment, the budget available was only sufficient to sequence the two chosen time points from each leaf section using the biological triplicates. Thus, the Q-RT-PCR analysis of individual genes of interest was extended to the full time course of 6 samples collected at 4 h intervals of the 12:12 LD cycle.

It was hoped that this more detailed time course would help to elucidate the timing of peak transcript levels for each gene in sufficient detail to allow connections to be made between the expression patterns of the TF genes and the known CAM genes such as PPC and PPCK. For example, if one of the novel TFs plays a direct role in binding to the promoter of PPC and driving its expression in the leaf tip and temporal regulation over the light dark cycle, then the transcript abundance of the TF would be expected to be phased ahead of the daily peak in PPC transcript levels in the leaf tip.

A secondary purpose of this experiment was to validate and/ or test the consistency of the transcript measurement results between the RNA-seq analysis and the Q-RT-PCR in terms of both the leaf development, proximal-distal (basal-tip) axis, as well as the light/dark regulation of transcript abundance in each leaf section. It is good practise in any high-throughput gene transcript quantification experiment, be it via microarrays or RNA-seq, to validate the findings from the transcriptome wide approach using Q-RT-PCR for a selection of differentially regulated genes. If the findings from the RNA-seq are faithfully reproduced using Q-RT-PCR, then this greatly strengthens the ability to rely on the RNA-seq data and draw reliable conclusions from the dataset.

As the CAM-associated CO_2 fixation carried out by the *A. sisalana* leaf tip was discovered to continue to oscillate robustly in LL conditions using gas exchange measurements (Figure 4.2), it was important to determine whether or not any of the discovered TF genes also oscillated with a robust circadian rhythm under LL as this would provide strong support for the TFs being candidates for playing a role in the circadian coordination of the biochemical steps of CAM and thus the CO_2 fixation rhythm. In order to test for circadian clock control of the TFs and the control CAM and white basal-enhanced gene(s), samples of the fully-developed-CAM leaf tip were collected under free-running LL conditions and used for total RNA isolation. These RNAs were also used for Q-RT-PCR analysis of the TF genes in order to determine which of the newly discovered CAM-induced and light/dark regulated TFs were subject to circadian clock control. This was important as although the LD data may suggest that a gene could be under circadian clock control, this can only be confirmed if the gene transcript abundance continues to cycle under constant conditions with no environmental inputs to the leaf.

The RNA samples used for the 24 h LD Q-RT-PCR measurements were the same ones used for RNA-seq in Chapter 5 and the same plant tissues were also used for metabolic measurement in Chapter 4 obtained and prepared as described in Section 2.1.2. For the LL leaf tip experiment, the samples were prepared as described in Section 2.1.3; these samples were also used for the analysis of circadian clock control of sugar and malate levels as described in Chapter 4. It was possible to make multiple different measurements using exactly the same biologically replicated leaf samples because *A. sisalana* leaves are relatively large, even for the youngest fully expanded leaf samples from the young 11-week-old plants that were used for the sampling of the different leaf sections and LD time points. Once the leaf samples had been ground to a fine powder in a pestle and mortar using liquid nitrogen, sub-aliquots could be used for the isolation of total RNA, soluble metabolites and the total leaf proteins used for the immuno-blot analyses. This means the results of these different measurements are directly comparable to one another, strengthening the validity of the cross-comparisons drawn in the discussion and conclusions.

The candidate CAM-associated regulatory TF genes analysed in this chapter included the novel CAM-induced genes (*AsAP2, AsBTB, AsHomeobox, AsPLATZ*, *AsNAC, AsWRKY,* and *As_zf_DOF*), the known positive control CAM-associated genes (*AsPPC* and *AsPPCK*), the non-CAM negative control gene (Class I *AsKNOX*), the control circadian clock gene (*CCA1)* and tip-induced control sugar metabolism gene *(As_cwINV*). *As_cwINV* also served as a control clock-controlled output pathway gene, as its transcript abundance level exhibited a robust circadian rhythm in the semi-quantitative RT-PCR results, performed previously and prior to the time when the Q-RT-PCR was available. The semi-quantitative RT-PCR result of *As_cwINV* was presented in this chapter (Figure 6.4D) along with the Q-RT-PCR results (Figure 6.4C).

6.2 Result and discussion

6.2.1 PCR efficiency and melting curve analysis

Primers for Q-RT-PCR were designed using the sequences of the novel CAM-induced TFs, CAM positive control genes, non-CAM negative control gene and other control genes for the circadian clock according to the criteria outlined in the Life Technologies™ Real Time PCR handbook (2014 edition). Primers were designed to the sequences of selected candidate genes extracted from the Illumina-Trinity assembly (Section 5.2.2) using the Geneious programme version 5.4 (Drummond *et al.*, 2011), as described in Section 2.4.1. The differential expression analysis was performed at Trinity 'gene' level (Section 5.2.3). Thus, one 'gene' contains one or more isoforms due to the fact that a Trinity 'gene' was a cluster of isoforms that shared sequence content (Grabherr *et al.*, 2011). Primers for Q-RT-PCR must be gene specific and must therefore be designed to only one particular sequence. Thus, multiple alignments of all isoforms of a gene were performed in order to identify the best aligned isoform that contained all the shared sequence content of all isoforms; this was usually the longest contig of each 'gene'. This particular isoform was then used for the Q-PCR primer design. To ensure the correct contig was obtained for the novel discovered CAM-induced genes, the sequence of selected isoform was searched using BLASTX against the rice protein database (http://rice.plantbiology.msu.edu), a monocot grass species that is more closely related to Agave than the model dicot C_3 species *A. thaliana*. The coding DNA sequence (CDS) of the top BLASTX hit which corresponded to the correct gene from rice database was then aligned with the longest ORF of the query Agave gene sequence. Consequently, all the Agave ORFs aligned and overlapped well with the rice CDSs, although some of Agave ORFs were shorter than rice CDSs due to the fact that the Agave ORFs were from the transcripts constructed from the sequence reads and did not always span the full length transcript (i.e. assembled *A. sisalana*

contigs in the Trinity assembly were often partial cDNAs). To determine the specificity of a contig (selected isoform for each gene) and primers, the longest ORF of that contig and the primers designed for that contig were searched back against the Agave assembly using Geneious. The longest ORF search result demonstrated the correct contig as the top hit followed by a few similar contigs which appeared to be other isoforms within the same gene and occasionally other paralogous genes. The primers search result returned only the contig which the primers were designed to, confirming the specificity of the primers. In addition, primers for circadian clock and reference genes were also designed to the sequences extracted from the Illumina assembly. These sequences were obtained from the corresponding sequences of the sequence search against Illumina assembly using the 454-assembled sequences previously used in designing primers for semi-quantitative RT-PCR experiment in Chapter 3 as search query sequences. These genes showed good results in semi-quantitative RT-PCR and should be good control circadian clock genes. *AsUBQ10* was used as the loading control reference gene for the Q-RT-PCR analysis as it had proven to be a good reference gene in the previous semi-quantitative RT-PCR work described in Chapter 3.

One of the essential quality controls of Q-RT-PCR is the PCR reaction efficiency (Bustin, 2010). Powerful and accurate Q-RT-PCR assays are usually correlated with the optimal PCR efficiency (Bustin *et al.*, 2009). According to Life Technologies™ Real Time PCR handbook (2014 edition), it is recommended that Q-RT-PCR amplification efficiency needs to be 90-110 %. In this study, all PCR reactions of all genes exhibited greater than 90 % PCR efficiency, placing the primers and amplification reactions within the optimal range (Supplementary Figure 6.1; red frame). However, ideally the PCR reaction efficiency should be 100% meaning that the amount of cDNA template doubles after each thermal cycle during the process of exponential amplification.

The specificity of a Q-RT-PCR assay can be assessed through the primers designed and reaction conditions applied. Nevertheless, even for well-designed primers, it is still possible that the primers might amplify a nonspecific product or primer-dimers. The specificity of the Q-RT-PCR reaction can be determined using post-amplification melting curve analysis (see method 2.4.2). The melting curve analysis is a very straightforward and easy approach to determine primer-dimer products in Q-RT-PCR reactions and assess the reaction specificity. Different PCR products obtained from the same reaction normally result in distinguishable melting characteristics (curves) due to the fact that the melting temperature of cDNA is influenced by GC content, length, and the base mismatches, along with other factors. Thus, PCR reaction products with uniform melting characterization in the same reaction can ensure the specificity of the reaction (Life Technologies™ Real Time PCR handbook, 2014 edition). The results presented here demonstrate that the melting curves of amplicons of all genes tested exhibit nearly identical melting characteristics of PCR products with single peaks (Supplementary Figure 6.2 and 6.3) implying that the amplified products are specific to the target cDNA.

6.2.2 Transcript abundance level of 24 h light/ dark time course

In general, comparison between the Q-RT-PCR results in this chapter (Figure 6.1 and 6.2) with the results derived from the RNA-seq DE analysis based on FPKM values in Section 5.2.4 (Figure 5.9 and 5.10), the transcript level results from the two different RNA quantification approaches for the same genes correlated very well in terms of both leaf development and light/dark regulation. However, there were some slight differences detected between the two sets of results. For example, *As_zf_DOF* (c534926_g1) transcript levels were higher in the white segment than in the leaf base at 22:00 according to the Q-RT-PCR results, while the leaf base was higher than the white segment in the RNA-seq FPKM analysis. Moreover, *AsPPCK* (c477309_g1) transcript was present in leaf base at 22:00 dark in the Q-RT-PCR analysis, while

this was nearly undetectable in the RNA-seq. This might be possibly due to the minor error during the procedures of either Q-RT-PCR or RNA-seq as the difference between the 2 results were relatively small and detected only at one time point for two of the studied genes. The results for the rest of the genes were correlated very well between Q-RT-PCR and RNA-seq FPKM values. In addition, the result of another contig of *AsPPCK* (c561463_g1) that was presented in the RNA-seq FPKM analysis (Figure 5.10E), was not presented here as the results of its transcript levels between Q-RT-PCR and RNA-seq analysis were highly different. This was due to the technical challenges, namely neither the first pair of primers designed nor a new pair of primers that were redesigned for *AsPPCK* (c561463_g1) worked with high percentage efficiency. It was also not possible to repeat the experiment any further as the cDNA samples ran out and was just enough to complete Q-RT-PCR amplifications of all the other genes. This was because plant tissues from the same samples were used in many experiments including semi-quantitative RT-PCR (chapter 3), physiological and biochemical analysis (chapter 4), and RNA-seq analysis (chapter 5), and the Q-RT-PCR presented in this chapter was the final experiment carried out.

The transcript abundance of all candidate genes based on FPKM values obtained through the RNA-seq analysis on different leaf segments sampled at 10:00 and 22:00 was already described and discussed in Section 5.2.4, Chapter 5. Here, the full 24 h LD time-course results are presented and discussed. *AsNAC* (c566713_g1), *AsWRKY* (c571790_*g2*) and *As_zf_DOF* (c534926_g1) demonstrated a very similar expression pattern of transcript abundance in the CAM-performing leaf tip samples (Figure 6.1A, B and F), all exhibiting a dark phased transcript peak. The leaf tip transcript level of *AsNAC* (c566713_g1) and *As_zf_DOF* (c534926_g1) started and remained at a low level during the light period, increased throughout the dark, and peaked at the end of the dark period (Figure 6.1A and F). *AsWRKY* (c571790_*g2*) also exhibited a similar pattern except for the fact that transcript level in the light did not go very low and the

difference of the *AsWRKY* (c571790_*g2*; Figure 6.1B) transcript abundance between light and dark samples was not as pronounced as it was for *AsNAC* (c566713_g1) and and *As_zf_DOF* (c534926_g1) (Figure 6.1A and B). In other leaf segments, the transcript level of these genes stayed very low throughout all time points, except in the case of *As_zf_DOF* (c534926_g1) where the transcript level rose at the end of the dark in a similar pattern to the leaf tip, but only reached about half of the tip abundance in both the pale green leaf base and white basal leaf tissue (Figure 6.1A, B and F). In RNA-seq analysis of the leaf longitudinal developmental gradient in the *Z. mays* (maize) leaf, these three TF families were expressed in different leaf segments. *zf_DOF* family genes were highly expressed in tip, while an approximately equal number of *NAC* family TFs were detected as being up-regulated in both base and tip, and *WRKY* family genes were up-regulated in the leaf base relative to the older section further up the leaf (Li *et al.*, 2010). In the *A. deserti* leaf, *WRKY* family was also abundant but did not show differential expression among the leaf segments while the *zf_DOF* and *NAC* families were not investigated or highlighted in the *A. deserti* proximal-distal leaf gradient study (Gross *et al.*, 2013). *AsPLATZ* (c541787_g1), *AsBTB* (c599899_g1) and *AsAP2* (c582092_g1) transcript levels were generally higher in the tip than the other leaf segments in this study (Figure 6.1C, D and E). These three TF families were not reported to be enriched in certain segments of the leaf developmental gradient in either *Z. mays* (maize) or *A. deserti* (Li *et al.*, 2010; Gross *et al.*, 2013). The difference in transcript level of *AsAP2* (c582092_g1) between leaf tip and other segments was not as pronounced as the difference for the *AsPLATZ* and *AsBTB* genes (Figure 6.1). *AsPLATZ* displayed one of the most striking results across all of the studied genes as it was virtually undetectable in both the pale green leaf base and the white basal tissue across all 6 time points, whereas it was abundant at all time points in the leaf tip, and displayed a strong light/ dark oscillation pattern with a peak at the beginning of the light and in the middle of the dark (Figure 6.1C). *AsAP2* did not display a clear pattern of light/ dark regulation either (Figure

6.1E). This particular contig of *AsAP2* was selected as a candidate gene due to its distinct expression pattern when compared to the other selected contigs and candidate genes (chapter 5). However, the several other *AsAP2* genes, detected based on the FPKM-based RNA-seq analysis but not selected as candidate genes, exhibited very strong differential transcript regulation in leaf development and light/dark regulation (data not shown). *AsBTB* (c599899_g1) transcript abundance in the leaf tip peaked at the beginning of the light period, reached its lowest level at the beginning of the dark, and followed by a steady rise throughout the dark period (Figure 6.1D). This pattern is very similar to that of *AsPPC* (c489202_g2; Figure 6.2B).

AsHomeobox (c526089_g4) exhibited light/ dark regulation in the leaf tip (Figure 6.2A), and its diurnal pattern showed an interesting degree of correlation with the diurnal cycle of sucrose concentration in the leaf tip (Figure 4.7C). It started at a low level at dawn, increased throughout the 12 h light period, peaked at the beginning of the dark period, and then declined throughout the dark period (Figure 6.2A). Homeobox family TFs have been reported to be highly abundant in the leaf base in maize (*Zea mays*), where they were proposed to perform various functions in leaf development (Li *et al.*, 2010).

Overall, this Q-RT-PCR analysis of the complete light/ dark cycle of time points for the three different leaf segments supports the RNA-seq findings which originally indicated that the newly discovered CAM-induced TF genes displayed a higher transcript level in the leaf tip relative to the other leaf segments at all time points throughout the 24 h LD period. This confirms that these genes are strong candidate CAM-induced transcription factors, as their transcript expression patterns are similar to CAM genes.

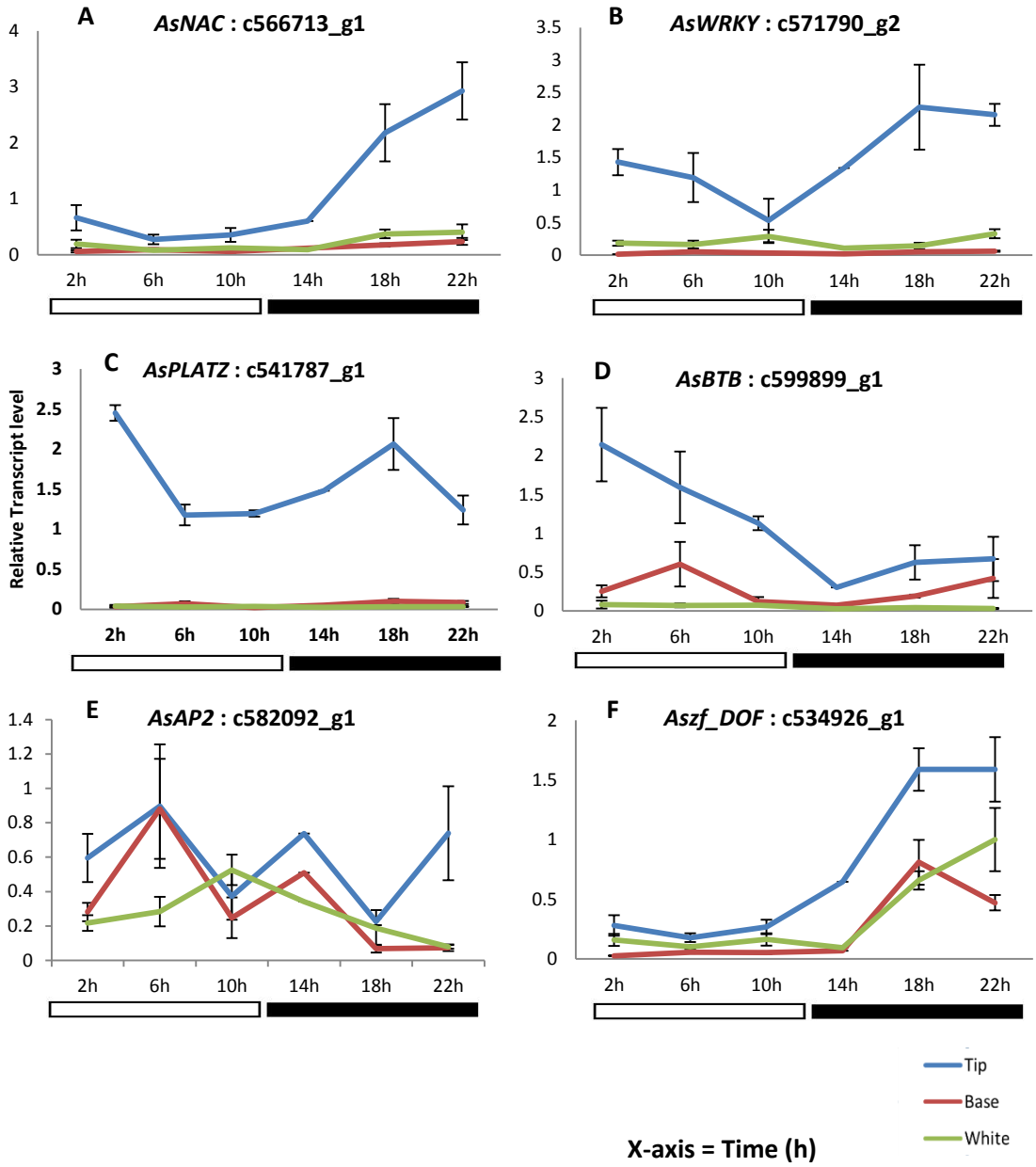


Figure 6.1 Q-RT-PCR analysis reveals the light/ dark pattern of transcript abundance regulation for six newly discovered CAM-induced transcription factor genes over a full 24 h light/ dark cycle contrasting different developmental leaf segments along the proximal-distal axis of the leaf.

Q-RT-PCR was used to measure the relative transcript abundance of transcription factor genes of interest in the youngest fully expanded *A. sisalana* leaf tip, base and white segment of 11-week-old plants sampled at 4 h intervals throughout a 12:12 light/dark cycle. The transcript values were normalized to the abundance of *UBQ10* transcripts. At the 14:00 sampling time point, only 1 biological replicate was calculated due to the missing of other 2 biological replicates caused by an accident with the frozen leaf tissue. The novel CAM-induced TF genes are as follows: *AsNAC* (A), *AsWRKY* (B), *AsPLATZ* (C), *AsBTB* (D), *AsAP2* (E), and *As_zf_DOF* (F). The Y-axis indicates the normalised relative transcript abundance. The X-axis indicates the time of sampling, with the white and black bars below each graph highlighting the light and dark periods respectively.

In terms of the non-CAM control gene, class I *AsKNOX1* (c568644_g2), its transcripts were only detected in the white basal section of the leaf (Figure 6.2D). This gene was not detected in the other sections of the leaf (Figure 6.2D). The gene therefore demonstrates very clearly that not all genes increase in the tip relative to the basal portions of the leaf, and thus that the tip-enhanced genes selected and studied in detail here are indeed strong candidates to play a functional role in the developmental induction and/ or light/ dark and circadian clock coordinated regulation of genes in the core CAM pathway. The class I *KNOX1* result was consistent with the other previous studies in *A. sisalana* (Zhou *et al.*, 2012) and in other species (Hake *et al.*, 2004; Hay and Tsiantis, 2009) where *knotted*-like homeobox (*Asknox*) transcript, determined using traditional Sanger EST sequencing and real-time PCR, was highly abundant in the apical meristem (positioned below the leaf base, possibly most comparable to the white basal leaf segment used here), but relatively low in other leaf tissues and undetectable in mature leaves.

The control CAM gene, *AsPPC* (c489202_g2), which encodes the main primary carboxylase that mediates CO_2 fixation in the dark period during CAM, showed very high transcript levels in the leaf tip compared to other leaf sections, as expected (Figure 6.2B). In addition, *AsPPC* transcript was also detected at a low level in the pale green leaf base samples, which was consistent with the presence of a faint PEPC protein band in the pale green leaf base protein samples (Section 4.2.2; Figure 4.3). This low level of PPC transcript and protein in the pale green leaf base tissue may be a sign that this C_3-performing section of the leaf is already preparing for CAM development. The transcript level was generally higher in the light than the dark samples. This result was again consistent with other previous results including those reported in Chapter 3, and published findings for obligate and facultative CAM *Clusia* species (Taybi *et al*., 2004). However, the dawn phased peak and dusk phased trough of *PPC* transcript level in the *A. sisalana* leaf tip contrasts strikingly with the pattern of light/ dark regulation of

PPC in CAM-induced *M. crystallinum* (Cushman *et al*., 2008), and in CAM leaves of *K. fedtschenkoi* (Dever *et al*., 2015), which both displayed a peak of CAM *PPC* transcript abundance phased to dusk, with a trough reached at dawn. In contrast, in *A. sisalana,* the transcript level of *AsPPC* peaked at dawn, declined through the light period, and rose steadily through the dark period (Figure 6.2B). Whilst the temporal phasing of *PPC* transcript abundance in *M. crystallinum* and *K. fedtschenkoi* is consistent with the time of day when the protein function is required (i.e. PPC transcripts peak at dusk and PPC activity is required throughout the dark period), the temporal pattern of *AsPPC* transcript abundance in *A. sisalana* would be more consistent with *AsPPC* functioning in the light as it peaks at dawn, so there are likely to be other levels of temporal control of PEPC beyond it being regulated by *de novo* transcription and translation.

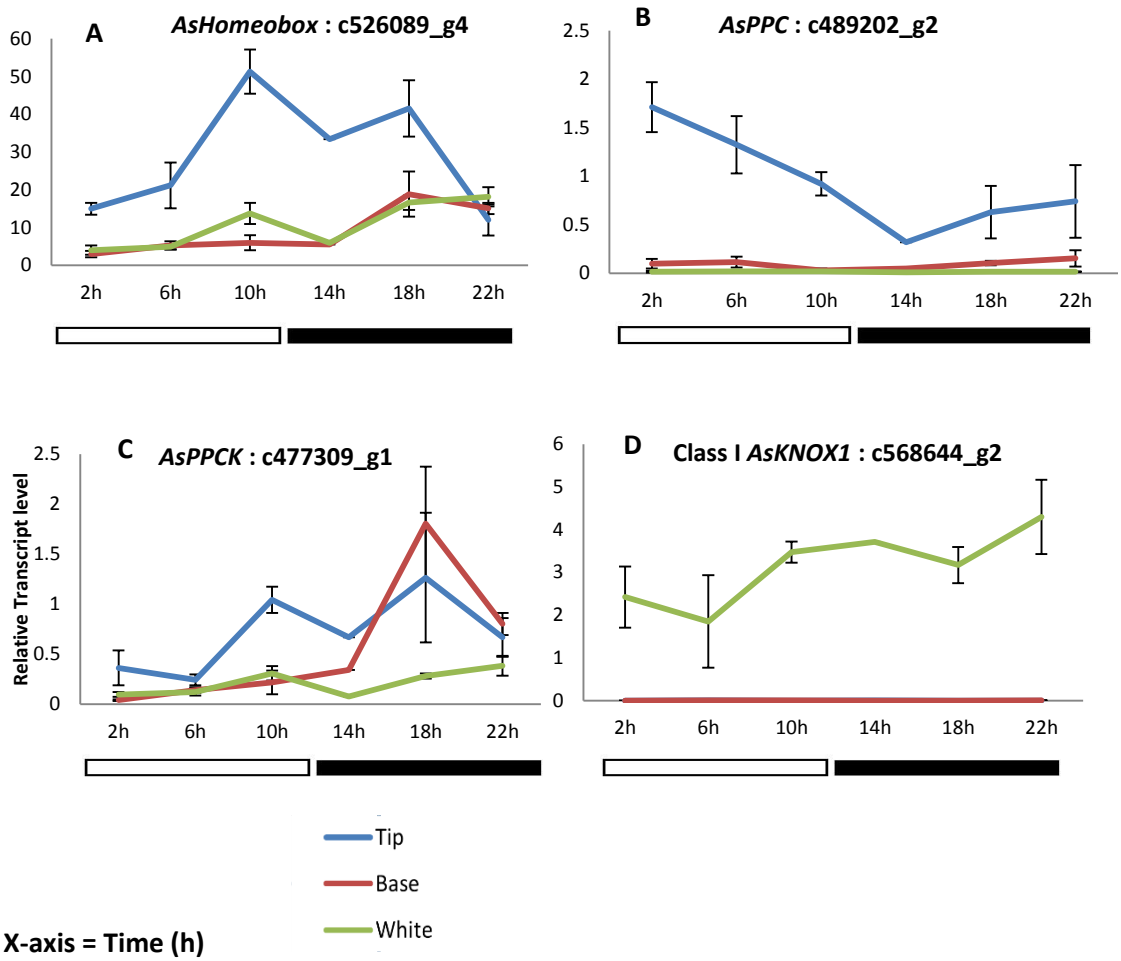


Figure 6.2 Q-RT-PCR analysis reveals the light/ dark pattern of transcript abundance regulation for a newly discovered CAM-induced transcription factor gene, known CAM control genes, and a non-CAM gene over a full 24 h light/ dark cycle contrasting different developmental leaf segments along the proximal-distal axis of the leaf.

Q-RT-PCR was used to measure the relative transcript abundance of transcription factor genes of interest in the youngest fully expanded *A. sisalana* leaf tip, base and white segment of 11-week-old plants sampled at 4 h intervals throughout a 12:12 light/dark cycle. The transcript values were normalized to the abundance of *UBQ10* transcripts. At the 14:00 sampling time point, only 1 biological replicate was calculated due to the missing of other 2 biological replicates caused by an accident with the frozen leaf tissue. The novel CAM-induced TF gene: *AsHomeobox* (A), known CAM control genes: *AsPPC* (B) and *AsPPCK* (C), and a non-CAM gene: class I *AsKNOX* (D). The Y-axis indicates the normalised relative transcript abundance. The X-axis indicates the time of sampling, with the white and black bars below each graph highlighting the light and dark periods respectively.

The CAM gene that was studied here as a positive control, *AsPPCK*, has already previously been discussed in Section 3.2.1 based on preliminary result using semi-quantitative RT-PCR on the 24 h LD time course samples (Figure 3.5B) and in Section 5.2.4 (Figure 5.10D) based on FPKM-value RNA-seq analysis using the 10:00 and 22:00 samples. Those results demonstrated unclear and on several occasions somewhat conflicting levels of transcript variation for *AsPPCK* among leaf segments and time points. In this chapter, the expanded 24 h LD time course results generated using Q-RT-PCR, which should provide a more accurate and reliable quantification of the transcript abundance of *AsPPCK*, still generated an unclear impression of the regulation of *AsPPCK* (c477309_g1) transcript levels along the proximal-distal leaf axis in *A. sisalana* over the light/ dark cycle (Figure 6.2C).

AsPPCK (c477309_g1) levels were lower throughout the light/ dark cycle in the white basal leaf tissue, which is consistent with a reduced requirement for PEPC phosphorylation in the leaf base section and also with the reduced level of *AsPPC* (c489202_g2) transcripts and PEPC protein abundance in the leaf base (Figure 6.2B). When comparing the pale green C_3 leaf base and the dark green CAM leaf tip with the results for the white basal leaf tissue, both green sections of the leaf displayed clear increases in *AsPPCK* (c477309_g1) transcript levels at a number of time points (Figure 6.2C). In particular, the pale green leaf base result was consistent the expected pattern of regulation of *AsPPCK* in a CAM leaf, with a clear peak of transcript abundance in the middle of the dark period (Figure 6.2C), which is when CAM leaves use PPCK to phosphorylate PEPC, making it less sensitive to feedback inhibition by malate throughout the dark period. However, the pale green leaf base was shown earlier, in chapter 4, to perform C_3 photosynthesis, fixing all of its carbon in the light period, and likewise the pale green leaf base had a low and steady level of malate throughout the light/ dark cycle. Thus, it was surprising to find that the *AsPPCK* (c477309_g1) gene was up-regulated in the pale green

base of the leaf when compared to the white basal tissue, especially considering that the up-regulation coincided with the dark period, which is commonly taken as a signature for CAM.

When comparing the transcript regulation in the dark green leaf tip section to the pale green base and white basal tissue, *AsPPCK* (c477309_g1) transcript levels were higher in the tip than the white tissue at all time points (Figure 6.2C), which supported the conclusion that *AsPPCK* levels were induced in the leaf tip that performs CAM relative to the basal white, non-photosynthetic tissue. However, the data for the CAM leaf tip did not show a classical pattern of a dark-phased increase in *AsPPCK* transcript abundance that was consistent with previous reports in other constitutive CAM species such as *K. fedtschenkoi* (e.g. Hartwell *et al*., 1999; Taybi *et al*., 2000). In the tip section of the leaf, *AsPPCK* (c477309_g1) was however lowest at the first two time points in the light (02:00 and 06:00), and rose in the late afternoon, remaining relatively high through to a peak in the middle of the dark period (Figure 6.2C). However, it needs to be noted that the error bars for the sample in the middle of the dark at 18:00 were large, thus the difference between leaf tip and base transcript needs to be judged with care at this time point (Figure 6.5C).

Overall, the *AsPPCK* (c477309_g1) transcript profile was still broadly consistent with the dark period increase in PEPC phosphorylation in the leaf tip that was characterised with immuno-blotting in chapter 4 (Figure 4.3), as the level of *AsPPCK* (c477309_g1) transcripts in the tip did rise late in the light period and remained high for the first half of the dark period (Figure 6.2C). The absence of PEPC phosphorylation in the light period in leaf tip does not necessarily imply that the transcript was not present, due to the possibility of post-transcriptional and post-translational control of the regulation of PPCK activity. The high level of *AsPPCK* (c477309_g1) transcripts in the pale green, C_3 leaf base samples in the dark may indicate that the leaf was already on a developmental trajectory towards the development of CAM in the young, pale

green C_3 leaf base. As already discussed in Section 5.2.4, another possible cause of the unclear result might be that *AsPPCK* in Agave might have different transcriptional expression from other CAM genes or from other species. In addition, there might be more than one *AsPPCK* gene with different patterns of regulation and roles in PEPC phosphorylation, similar to the scenario reported previously for *A. thaliana* by Fontaine *et al.*, (2002) where two *PPCK* genes were differentially expressed in different parts of the plant, and with different responses to light signalling.

Noticeably, another contig (*c528974_g6*; annotated as *PPCK*), based on FPKM values from RNA-seq analysis in chapter 5, was found to be highly expressed in white samples, while other known CAM genes were only found highly expressed in leaf tip and not in other segments (data not shown). When aligned with the tip-highly-expressed *AsPPCK* (c477309_g1) presented here, based on FPKM values from RNA-seq analysis, the alignment of the two contigs indicated only 53.8 % pairwise identify and 16.3% identical sites. The primers used here did not align to this white-basal tissue up-regulated *PPCK* contig. This indicates that these two *PPCK* contigs are likely to encode paralogs with different functions. Moreover, *AsPPCK* transcript abundance (maximum FPKM values of 150) was noticeably lower than the other known CAM gene contigs, which exhibited much higher FPKM values, especially *PPC* with maximum FPKM of more than 3,000 (data not shown). This is consistent with the low abundance of *PPCK* transcript and protein in *K. fedtschenkoi* (Hartwell *et al.*, 1999).

These inconsistent patterns of *AsPPCK* regulation in *A. sisalana* leaves when compared to other CAM genes remain an area for further more detailed investigation in the future. Overall, taking into consideration the results for the two known positive control CAM genes studied in this detailed 12:12 LD experiment, the different transcript expression patterns of *PPC* and *PPCK* between *K. fedtschenkoi* and *A. sisalana* might be the consequence of the fact that

Agave plants evolved CAM completely independently from species such as *K. fedtschenkoi*. Therefore, they might have certain CAM genes with different functions that could be reflected by different expression patterns.

6.2.3 Q-RT-PCR analysis of the circadian regulation of the novel CAM-induced TF genes under constant light free-running conditions

It has long been known that certain CAM pathway genes are under circadian clock control, and that key features of the CAM system, including dark period phased CO_2 fixation, exhibit endogenous circadian rhythmicity under constant conditions (Hartwell *et al.*, 1996; Hartwell *et al.*, 1999; Nimmo, 2000; Hartwell, 2005; Mallona *et al.*, 2011). The results from the constant light, temperature and humidity (LL) full-CAM leaf tip time course gas exchange experiment in Chapter 4 also showed a robust circadian oscillation of CO_2 fixation and sugar levels, and weak rhythmicity of malate levels (Figs. 4.2, 4.6 and 4.8). For the results presented in the following section, a selection of the genes that showed a strong 24 h light/ dark pattern of regulation from amongst all of the TF candidates discovered via the RNA-seq analysis were selected for Q-RT-PCR using the *A. sisalana* leaf tip (fully developed CAM) total RNA samples collected every 4 h over an 82 h period under LL experimental conditions. These experiments allowed the identification of the CAM-induced TFs that were subject to persistent circadian oscillations in their transcript abundance in the absence of environmental inputs to the leaves. A circadian clock gene (*AsCCA1*) and sucrose metabolism associated gene (*As_cwINV*), which had previously shown a robust circadian rhythm when assayed with semi-quantitative RT-PCR analysis, were used as circadian clock-controlled positive control genes to test the validity of the results.

AsNAC (c566713_g1), *AsWRKY* (c571790_*g2*), and *AsPLATZ* (c541787_g1) all showed robust circadian oscillations throughout the LL conditions (Figure 6.3A, B and C). They also exhibited a similar pattern to the subjective light/ dark regulation of their transcript level, as their transcript peaks were phased initially to the end of the subjective dark and 2 h after subjective dawn. The pattern of subjective light/dark regulation of these genes correlated well with their results for the 24 h LD experiment (Figure 6.1A, B and C). This rhythm persisted throughout the entire LL time course, indicating that these CAM-induced transcription factors are circadian clock controlled and thus that they are good candidates for transcription factors that couple the central circadian clock to CAM. It is noteworthy that the timing of the peak transcript levels of these TF genes on each successive cycle happened earlier relative to the original light/ dark cycle under which the plants were entrained such that by the third peak, each gene peaked at the beginning of the dark period (62 h LL) whereas they all peaked at the end of the dark period or beginning of the light period during their first LL peak (at either 22 h or 26 h LL; Figure 6.3).

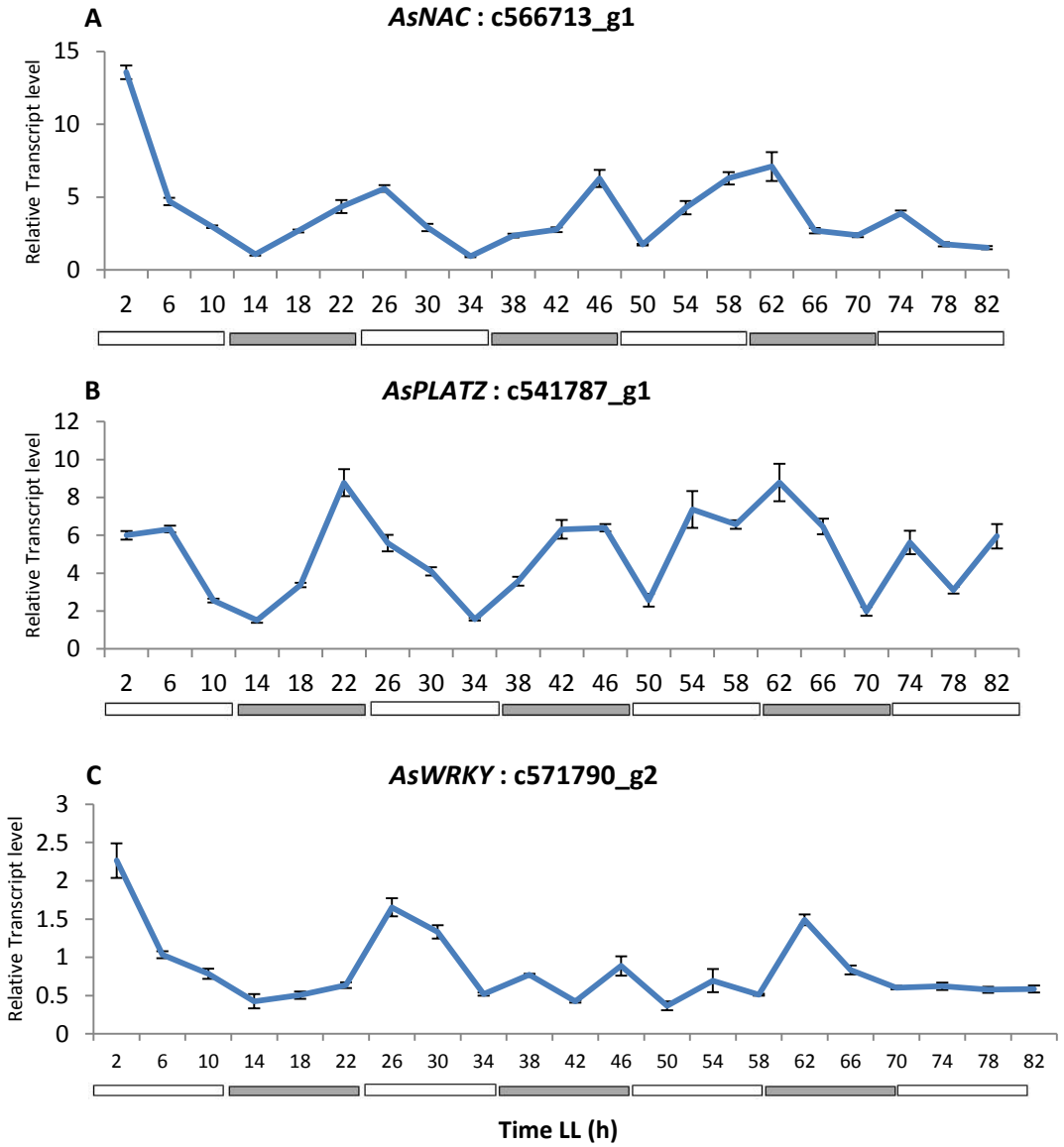


Figure 6.3 Q-RT-PCR analysis reveals oscillation of transcript abundance regulation for three newly discovered CAM-induced transcription factor genes over in constant light and temperature conditions in the leaf tip of *A. sisalana.*

Q-RT-PCR was used to measure the relative transcript abundance of transcription factor genes of interest in the youngest fully expanded *A. sisalana* dark green leaf tip of 11-week-old plants sampled at 4 h intervals throughout free-running constant light (LL) conditions for 82 h. The transcript values were normalized to the abundance of *UBQ10* transcripts. The novel CAM-induced TF genes are as follows: *AsNAC* (A), *AsWRKY* (B), *AsPLATZ* (C). Three technical replicates were used instead of biological replicates due to technical limitation of a number of similar sized plants available at the time of experiment, plus the limited room in the Snijders Microclima MC-1000 growth cabinet that could not accommodate 63 plants at the same time. The Y-axis indicates the normalised relative transcript abundance. The X-axis indicates the time of sampling, with the white and grey bars below each graph highlighting the subjective light and dark periods respectively.

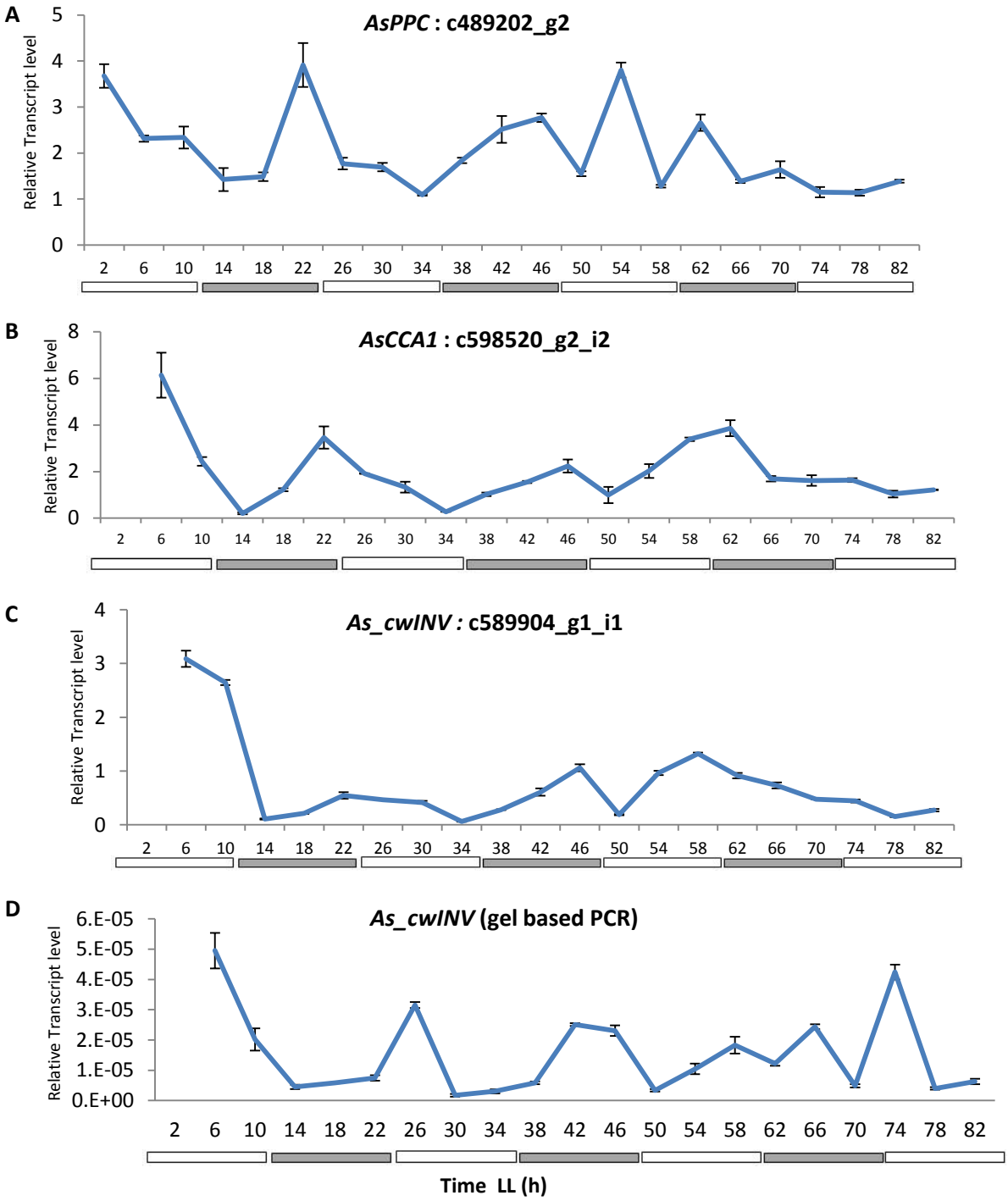


Figure 6.4 Q-RT-PCR analysis reveals oscillation of transcript abundance regulation for control CAM and circadian clock genes over in constant light and temperature conditions in the leaf tip of *A. sisalana.*

Q-RT-PCR was used to measure the relative transcript abundance of transcription factor genes of interest in the youngest fully expanded *A. sisalana* dark green leaf tip of 11-week-old plants sampled at 4 h intervals throughout free-running constant light (LL) conditions for 82 h. The transcript values were normalized to the abundance of *AsUBQ10* transcripts. A control CAM gene: *AsPPC* (A). The value of 2 h time point was not plotted in the graph of circadian clock-controlled positive control genes: *AsCCA1* (B; 10.03), *As_cwINV* (C; 15) and *As_cwINV* (Gel base; D; 1.09E-04) as it contained exceedingly high

transcript value which would make the graph look flat and difficult to determine the circadian rhythm. *As_cwINV* (gel based PCR; D) was obtained using semi-quantitative RT-PCR previously performed when Q-RT-PCR was not available. Three technical replicates were used instead of biological replicates due to technical limitation of a number of similar sized plants available at the time of experiment, plus the limited room in the Snijders Microclima MC-1000 growth cabinet that could not accommodate 63 plants at the same time. The Y-axis indicates the normalised relative transcript abundance. The X-axis indicates the time of sampling, with the white and grey bars below each graph highlighting the subjective light and dark periods respectively.

The control CAM gene transcript, *AsPPC* (c489202_g2), also showed a circadian rhythm with constant peaks at the first two subjective dawns under LL conditions (Figure 6.4A). The central circadian clock control gene, *CCA1*, is known to peak at dawn and show robust circadian rhythmicity (Wang and Tobin, 1998; Park, 1999). In this study, the *AsCCA1* (c598520_g2_i2) gene that was assayed had a pattern of regulation that was highly consistent with its known free-running rhythm in species such as *A. thaliana* and *K. fedtschenkoi* (Figure 6.4B). This supports the proposed idea that genes phased to the same time of the day as *AsCCA1* are oscillating with a reasonably robust rhythm. *As_cwINV* (c589904_g1_i1), a sucrose metabolism related gene, showed a robustly rhythmic oscillation with both the Q-RT-PCR analysis shown here and the semi-quantitative RT-PCR result, although the semi-quantitative RT-PCR result showed a more robust more robust oscillation (Figure 6.4C and D). Overall, the control circadian clock controlled genes demonstrated good results with robust rhythmicity in free-running conditions.

6.3 Summary

The results of the transcript abundance determination using Q-RT-PCR on the leaf samples obtained from the 24 h LD and LL free-running time course experiment (Figure 6.4, 6.5 and 6.6) have generally worked well. The results provide greater details defining the transcript abundance profiles of the novel CAM-induced transcription factors discovered via the *de novo* RNA-seq analysis in Chapter 5 which only relied on two sampling time points (light: 2 h before

dusk and dark: 2 h before dawn). From the results of the 24 LD experiment, it was possible to interpret which of the newly discovered CAM-induced transcription factors exhibited strong light/ dark regulation and would be likely to robustly oscillate under the constant free-running conditions. The Q-RT-PCR was also aimed to validate the RNA-seq analysis results, determining if the two results are identical in terms of the transcript levels of the same samples and time points.

The results for 24 LD time course were very well correlated with the RNA-seq data based expressed as FPKM values (Figure 5.9 and 5.10). This was the case both in terms of leaf development and light/ dark regulation. Among the novel discovered CAM-induced genes, *AsNAC* (c566713_g1), *AsWRKY* (c571790_*g2*) and *As_zf_DOF* (c534926_g1) showed a very similar expression pattern in the leaf tip where they were higher in the dark than the light, peaking at the end of the dark period (Figure 6.1A, B and F). In the *Z. mays* (maize) leaf, these three transcription factor families were also present and highly expressed in different leaf segments in which the regulation patterns differed greatly from the results of the same transcription factors presented here (Li *et al.,* 2010). *zf_DOF* family genes were up-regulated in tip, while many *NAC* family TFs were highly expressed in both base and tip, and TFs of *WRKY* family were up-regulated in the leaf base (Li *et al.,* 2010). The result in terms of light/ dark regulation was not given (Li *et al.,* 2010). It was perfectly possible that in different species, the transcription factors from the same family would have different regulation patterns as they would have totally different functions in different leaf tissues. In the *A. deserti* leaf study, among the three families of transcription factors presented here, only *WKRY* was highlighted, but no information was given as to how it was differentially expressed (Gross *et al.,* 2013). It would be ideal if it was possible to compare the regulation patterns of transcript levels of the same transcription factors of the closely related Agave species. The transcript abundance of *AsPLATZ* (c541787_g1), *AsAP2* (c582092_g1) and *AsBTB* (c599899_g1) was generally higher in

the tip than the other leaf segments throughout the LD time course (Figure 6.1C, D and E). However, these three transcription factor families were not highlighted as being potentially important in either the study on *Z. mays* (maize) or the *A. deserti* proximal-distal leaf developmental gradient (Li *et al.*, 2010; Gross *et al.*, 2013). *AsBTB* (c599899_g1) showed a very interesting transcript expression pattern, which was very similar to that of the control CAM gene, *AsPPC* (c489202_g2; Figure 6.2C). It peaked at dawn, declined throughout the light period and rose steadily throughout the dark period. The LD transcript abundance profile of the gene *AsHomeobox* (c526089_g4) displayed a pattern of regulation over the light/ dark cycle that was similar to that of the sucrose concentration in the leaf tip (Figure 4.7C). Class I *AsKNOX1* (c568644_g2) was determined to be a good non-CAM negative control gene, as it was detected solely in the white basal leaf section. This was consistent with the regulation of related *AsKNOX* genes in previous studies in *A. sisalana* (Zhou *et al.*, 2012) and other species (Hake *et al.*, 2004; Hay and Tsiantis, 2009), as Class I *KNOX* genes are involved in the formation of shoot apical meristem (SAM) during embryogenesis, and the function of SAM throughout the whole lifetime of the plant (Scofield *et al.*, 2013), which normally occurs in the youngest developing part of the leaf (white basal segment).

The control CAM gene, *AsPPC* (c489202_g2), displayed a higher transcript level at the beginning of the light than in the dark samples (Figure 6.2B), which correlated well with the semi-quantitative RT-PCR work presented in Chapter 3. The other control CAM gene, *AsPPCK* (c477309_g1) exhibited somewhat unexpected results compared to the known pattern of regulation of this CAM gene in other CAM species (Figure 6.2C). In particular, the difference in transcript level between the CAM leaf tip and the C_3 base was not very great unlike previous results in other CAM species which have shown that PPCK is induced concomitant with CAM and comes under strong light/ dark and circadian clock control in the CAM performing leaves (Hartwell *et al*., 1999). *PPCK* in Agave might have different transcript expression characteristics

from other CAM genes or from other species. Another *AsPPCK* contig (c528974_g6) was found to be an up-regulated transcript in the white basal leaf samples based on FPKM values from RNA-seq analysis (data not shown), whilst the other *AsPPCK* contigs were highly expressed in leaf tip (Figure 5.10 D and E). This distinct *AsPPCK* expression reveals the existence of a paralogous *AsPPCK* gene, which likely plays a distinctive role in the regulation of PEPC in the leaf basal region. However, these unexpected *AsPPCK* results merit further investigation. In particular, it will be important to investigate whether or not there are other *AsPPCK* paralogs in *A. sisalana,* and if so, what their pattern of regulation is. In terms of light/ dark regulation, *AsPPCK* transcript levels appeared to be higher in dark than the light samples in both leaf tip and pale green base, which correlated with the previous studies in *K. fedtschenkoi* and *M. crystallinum* (Hartwell *et al*., 1999; Taybi *et al*., 2000; Boxall *et al*., 2005).

The results for *AsPPC* transcript*,* PEPC protein, and *AsPPCK* transcript abundance in *A. sisalana* may indicate early preparation for CAM in the young C_3 leaf base that will finally become CAM-induced as the leaf grows and matures. In addition, the light/ dark regulation patterns of *AsPPC* and *AsPPCK* transcript were found to be different between *K. fedtschenkoi* and *A. sisalana*. This might be due to the fact that Agave evolved CAM completely independently from Kalanchoe species, resulting in CAM genes with different functions and expression patterns.

In the Q-RT-PCR analysis of the LL free-running time course experiment, *AsNAC* (c566713_g1), *AsWRKY* (c571790_*g2*), and *AsPLATZ* (c541787_g1), selected as being transcription factors that demonstrated the most robust light/ dark regulation, showed similar subjective light/dark regulation (peaking at late night or dawn) and a robust circadian rhythm confirming that they are circadian clock controlled (Figure 6.3A, B and C). *AsPPC* (c489202_g2), a control CAM gene, also exhibited a clear circadian rhythm with peaks at subjective dawn (Figure 6.4A). The

control clock gene *AsCCA1* (c598520_g2_i2; Figure 6.4B) and clock-controlled output pathway gene *As_cwINV* (c589904_g1_i1), the latter measured using both Q-RT-PCR and semi-quantitative RT-PCR (Figure 6.4C and D) also showed robust oscillations of their transcript abundance under LL conditions in the leaf tip. The results for these control genes support the conclusion that the Q-RT-PCR method, primers and samples applied were reliable for the identification of clock controlled genes in the *A. sisalana* leaf tip region.

Overall, this Q-RT-PCR analysis performed as a follow-up to the RNA-seq analysis has produced promising results, which are a valuable addition to this study. Following the identification of the new transcription factors from the RNA-seq work, the Q-RT-PCR results demonstrated that these newly discovered genes cycled robustly in both LD and also in LL conditions, confirming that they are CAM-induced and clock-controlled. Thus, they are good candidates for transcription factors that couple the central circadian clock to CAM. Due to the limited time, only certain transcription factors were identified and investigated. It would be worthy to continue identifying and testing more transcription factors if there were further funding and future possibilities for more detailed experiments.

Furthermore, it would extremely interesting to generate transgenic lines of *A. sisalana* in which the discovered genes were down-regulated with either RNAi, or a CRISPR/ Cas9 targeted gene disruption mutagenesis approach. However, early attempts to establish *A. sisalana* in tissue culture, aimed at developing a regeneration system that could underpin the development of a stable transformation system did not go well, with all tissue cultures succumbing to infection and being lost. Considering the relatively slow developmental rate of *A. sisalana*, certainly relative to a rapid cycling annual species such as *A. thaliana*, the development of a transformation system for *A. sisalana* was deemed to be beyond the scope of this PhD. If stable transgenic lines of *A. sisalana,* in which the novel genes identified here

were down-regulated with RNAi, or mutated with a system such as CRISPR-Cas9 gene editing, could be generated in the future, detailed phenotypic characterization of CAM in the transgenic lines that lacked the function of the identified genes in *A. sisalana* plants could be undertaken in order to discover whether or not the genes perform a function in the developmental induction and/ or light/ dark and circadian regulation of CAM in *A. sisalana.* The experiments to study the phenotypes of the transgenic lines could be performed to measure growth of the plants, gas exchange and CAM-related metabolites (e.g. malate and sugars) under light/ dark and free-running LL conditions, the abundance of key CAM-related proteins (e.g. PEPC, PPDK and their phosphorylated forms) and their activity. This way, it would be possible to determine if the transcription factor genes discovered here have a direct or perhaps indirect impact on the CAM system of *A. sisalana.*

Chapter 7

General Discussion

An investigation of the timing and localization of peak transcript abundance for a range of known CAM-associated genes in *A. sisalana*

This PhD thesis has presented and discussed results of a series of experiments using various techniques aimed to explore the biochemistry and functional genomics of CAM in *A. sisalana* with the key goal to achieve a transcriptome-wide view of the genes that *A. sisalana* employs to perform CAM. Preliminary experiments were performed using semi-quantitative RT-PCR in order to determine the transcript abundance of key CAM-related genes in *A. sisalana* plants of different ages, and proximal-distal leaf sections of different developmental stages, in order to distinguish between C_3 and CAM leaf tissues in *A. sisalana*. The results of these experiments provided a valuable guide for the appropriate selection of samples for the main RNA-seq experiment, which formed the central and major package of work within this project. In general, the results of the semi-quantitative RT-PCR experiments suffered from relatively high levels of variability, most probably as a result of the lack of biological replicates, which was an unfortunate consequence of the lack of available individual *A. sisalana* plants early in the project due to supplier nursery only being able to supply a small number of plants. It proved relatively difficult to define clear patterns of transcript abundance for the studied positive control CAM-associated genes, including *AsPPC, AsPPDK* and *AsPPCK,* in relation to both leaf development and light/ dark regulation (Chapter 3). This was in particular the case for *AsPPCK* regulation relative to the known regulation of this CAM-induced, clock-controlled gene in well-studied dicot model CAM species such as *K. fedtschenkoi* and *M. crystallinum* (Hartwell *et al*.,

1999; Taybi *et al*., 2000). However, it should be noted that *AsPPCK* regulation has not been reported or studied previously for a monocot CAM species, and thus, the expected results could only be predicted based on published work for dicot species. As CAM has evolved independently in many monocot and dicot lineages, it is perhaps not surprising that the regulation of *AsPPCK* detected here for *A. sisalana* was not as clear cut as the regulatory control reported for the dicot species.

However, the regulation of *AsPPC* and *AsPPDK* was, in broad terms, consistent with the previous studies; their transcript levels were mostly higher in the dark green tip relative to the pale green leaf base and white basal part of the youngest fully expanded leaf of *A. sisalana* (Figure 3.4A and B; Figure 3.5A and C). Despite the fact that the transcript abundance of *AsPPC* and *AsPPDK* was not always strikingly higher in the leaf tip samples and time points relative to the pale green base, the transcript abundance in the leaf tip was always clearly higher than the white basal section of the leaf (Figure 3.6A and B). *AsGI* was also included in the experiment as a circadian clock positive control gene (Figure 3.6C), and *As_cwINV* was included as a positive control as it demonstrated the clearest result in terms of both leaf developmental up-regulation in the tip, and light/ dark regulation in the tip where it peaked at dawn (Figure 3.6D). Based on these preliminary results for *AsPPC, AsPPDK, AsGI* and *As_cwINV* from the preliminarily scoping studies (Figure 3.5; Figure 3.6), the difference in transcript levels between C_3 leaf base tissues and the CAM leaf tip samples provided strong supportive evidence to guide the selection of samples for the RNA-seq experiment.

Physiological and biochemical analysis of CAM-associated characteristics

In addition to the semi-quantitative RT-PCR experiments in chapter 3, the metabolic and physiological analysis of CAM in *A. sisalana* presented in chapter 4 was designed to investigate

the level of CAM induction during leaf development, and its light/ dark regulation. Gas exchange analysis of leaf section CO_2 fixation demonstrated very clearly that the *A. sisalana* leaf base fixed its CO_2 solely during the light period, while the dark green leaf tip exhibited a full CAM CO_2 fixation pattern, with all four CAM phases present (Figure 4.1). The two transitional middle sections of the leaf also displayed some ability to fix atmospheric CO_2 in the dark period indicative of weak CAM, but also displayed pronounced CO_2 fixation in the light period. Furthermore, the dark green leaf tip from a mature leaf displayed strong/ full CAM under 12:12 light/ dark conditions followed by robust, high amplitude circadian rhythm of CO_2 fixation under LL free-running conditions (Figure 4.2).

Immuno-blot experiments which tested for the abundance of several key CAM-associated proteins demonstrated that key CAM enzymes were present nearly exclusively in the leaf tip, with a very low level of PEPC and β NAD-ME present in leaf base (Figure 4.3), indicating the nearly exclusive presence of critical CAM proteins in the leaf tip. PEPC phosphorylation (a direct result of PPCK activity) exhibited a striking and strong pattern of leaf tip induction and light/ dark regulation, with a high level of phospho-PEPC detected in the dark while the phospho-PEPC signal was not detected at all in the light samples of leaf tip, or in the basal samples from either pale green or white tissue. On the other hand, the other CAM enzymes did not show a pronounced light/ dark pattern of regulation on the immune-blots (Figure 4.4). Malate assays indicated that the light/ dark pattern of leaf malate concentration in the leaf tip was consistent with a typical daily malic acid pattern that has been well-studied and widelty reported for many CAM species (Black and Osmond, 2003). Malic acid accumulates in the vacuole during the dark period due to CO_2 fixation by PEPC, and the declines during the following light period due to malate leaving the vacuole and being decarboxylated by malic enzyme releasing CO_2 for secondary refixation by RuBisCo. In the leaf basal sections, there was no diurnal rhythm of leaf malate content, and malate levels were stable around the level of

the daily minimum towards the end of the light in the leaf tip (Figure 4.5). These results further strengthen the evidence for CAM induction only occurring in the leaf tip. However, the malate level in the leaf tip only showed a fairly weak circadian rhythm when assayed under LL free-running conditions (Figure 4.6). By contrast, soluble sugar assays revealed that sucrose exhibited a robust circadian rhythm throughout the LL time course, which ran for 82 h (Figure 4.8). The level of sucrose was considerably higher in the leaf tip than in the white basal sections, and almost undetectable in leaf base (Figure 4.7C). In the leaf tip, the sucrose level showed a clear light/ dark pattern of regulation with a peak of sucrose phased to dusk (Figure 4.7C), which was consistent with a typical daily pattern of reciprocal carbohydrate/ malate turnover associated with the CAM pathway (Black and Osmond, 2003). Fructose levels were significantly higher in the leaf base, which was consistent with previous reports that fructans were synthesised and stored in the leaf base in the grass *L. perenne* (Morvan-Bertrand *et al.*, 1999). The leaf glucose level was highest in the leaf base and the white basal section of the leaf, and relatively low in the tip. However, fructose and glucose generally did not exhibit a light/ dark pattern of regulation, except a very weak light/ dark pattern for the leaf tip of glucose level (Figure 4.7A and B). The results from the LL experiment showed a robust oscillation in sucrose and glucose levels, whilst fructose levels did not display a clear rhythm (Figure 4.8). Based on the metabolic and physiological results obtained and discussed, significant further support was generated to support the original hypothesis that the development of CAM in *A. sisalana* progressed from non-photosynthetic tissue in the most proximal white basal section, through C_3 in the pale green tissue at the leaf base, to full CAM in the distal leaf tip. This greatly assisted the selection of the most informative plant age and leaf developmental stage and sections, and time points, for the RNA-seq experiment, and also supported the comparative analysis of the RNA-seq results, with the biochemical and physiological data for the same leaf sections.

Comprehensive RNA-seq analysis of the genes associated with CAM and a detailed Q-RT-PCR investigation into the light/ dark and circadian clock control of CAM-induced transcription factors identified using RNA-seq

The comprehensive RNA-seq data analysis of the genes involved in CAM in *A. sisalana* leaf development and light/ dark regulation was performed using samples selected based on the results of previous experiments described above. Each light and dark sampling was carried out only at a single time point, 10:00 light (2 h before dusk), and 22:00 light (2 h before dawn), due to the available consumable budget to fund the sequencing library production and Illumina Hi-Seq sequencing runs. The transcriptome assembled using the Trinity pipeline from the resulting RNA-seq reads contained 671,886 Trinity 'genes', and 941,989 isoforms, which encoded 118,807 protein-coding sequences. In general, the analysis of transcript levels across different leaf segments from base to tip revealed biological processes and core classes of genes including transcription factors that were in common with those reported in *Z. mays* (Li *et al.*, 2010), and *A. deserti* (Gross *et al.*, 2013). This indicates a general evolutionary conservation of genes involved in leaf development in monocotyledonous model plants (Freeling, 1992). Due to the known role of transcriptions factors (TFs) in the developmental, light/ dark and circadian clock control of many biological processes in plants, candidate CAM-associated TFs were identified from within the list of differentially expressed genes from the RNA-seq analysis. TFs were selected based on their annotation within the 'Trinity' genes annotation file and their CAM-induction pattern (higher in leaf tip than other segments), and light/ dark regulation (Table 5.3). The proposed-CAM-candidate TFs were studied further using Q-RT-PCR in order to determine their transcript levels in relation to light/ dark regulation over the 24 h 12:12 LD time course. The TFs were also measured through Q-RT-PCR using the LL timecourse of leaf tip RNA samples. This analysis generated a number of interesting and

promising results which are a worthy addition to the RNA-seq analysis. Newly identified TFs included *AsNAC* (c566713_g1), *AsWRKY* (c571790_*g2*), and *AsPLATZ* (c541787_g1), which demonstrated a robust cycle of their transcript abundance under 12:12 LD conditions, and similar subjective light/ dark regulation (peaking at late night or dawn) under LL, with a robust circadian rhythm throughout the LL conditions (Figure 6.3A, B and C). This demonstrated that these selected TFs were CAM-induced, and clock-controlled, and thus that they were good candidates for TFs that may couple the central circadian clock to CAM.

***In silico* prediction of the CAM pathway in the mature leaf tip in *A. sisalana* based on RNA-seq derived transcript abundance patterns for the associated genes.**

Based on the quantitative transcriptome data generated from the RNA-seq data analysis, and previously known findings about the CAM pathway, it was possible to predict a novel model for the CAM pathway in the *A. sisalana* leaf tip for the first time, providing a testable framework of genes and proteins required for each step of the CAM pathway (**Error! Reference source not found.**). This diagram of the newly proposed pathway reveals how CAM may work in *A. sisalana* (**Error! Reference source not found.**), and can be supported by changes in the transcript levels of associated genes in the RNA-seq data. The pathway was modified from the simple proposed CAM pathway described in Figure 1.1 (Borland *et al.*, 2009) with more potential enzymes involved in CAM, sugars and fructan metabolism added. These genes were obtained from differentially expressed genes analysed from RNA-seq data. FPKM values for the most CAM-like gene transcript profile for each gene family were plotted and added into the pathway to highlight the transcript abundance pattern associated with leaf development and light/ dark regulation. As mentioned previously for the method used for the selection of the candidate TFs described in Section 5.2.4, genes with FPKM values higher than 10 were chosen. Genes showing CAM pattern in transcript level (FPKM; higher in leaf tip) were

also considered as the best candidates for CAM-related genes. To obtain the best possible representative of each gene presented in the pathway, *in silico* analysis was performed. The annotation information of all genes previously generated using Trinotate was used as a guide to sort which contigs represented which genes. In addition, an additional search was performed to identify the published, well-defined sequences for each gene by using BLAST searching against the GeneBank (Benson *et al.*, (2013); http://blast.ncbi.nlm.nih.gov/Blast.cgi) and Arabidopsis databases (Lamesch *et al.*, (2011); https://www.arabidopsis.org/Blast/). The protein sequence of the well-defined gene was used to BLAST search for the best candidate contig in the *A. sisalana* Illumina transcriptome using Geneious (Drummond *et al.*, 2011). The best candidate contig sequence was then BLAST searched back to the GeneBank database to ensure that it hits the correct genes in the GeneBank database, and that it is the correct contig for the searched gene. This process helped to ensure the correct identification of the *A. sisalana* contigs that encoded each potential CAM enzymes, transporters or regulatory proteins.

The physiology of CAM pathway was already and described in detail earlier in Section 1.1.2 (Figure 1.1). Here, the newly proposed CAM pathway was modified and filled with more CAM enzymes and details. Additionally, sugar and particularly fructan-related enzymes were also added in the pathway (**Error! Reference source not found.**).

The nocturnal fixation of CO_2 by PEPC and associated accumulation of malic acid in the vacuole involves several enzymes, which can collectively be referred to as the “carboxylation pathway” (Borland *et al.*, 2014). Among the CO_2 fixation-related enzymes, the transcript level of genes encoding β-carbonic anhydrase (*Asβ-CA*), *AsPPC* and *AsPPCK* exhibited clear CAM induction, i.e. up-regulated in the leaf tip relative to the basal sections of the leaf (Figure 7.1). Light/ dark regulation was not evident for *AsPPC* and *Asβ-CA* transcript levels, but *AsPPCK* showed a

higher transcript level for the leaf tip light samples, although the associated error bar was large (Figure 7.1). The transcript level regulation of *AsPPCK* in *A. sisalana* was found to contrast with expectations based on previously published work in other CAM species (more detailed discussion in Section 3.2.1, 5.2.4, and 6.2.2). The transcript levels of genes encoding enzymes involved in malate synthesis and transport (cytosolic *AsMDH*; *AsALMT*), and the H^+ transport required for the nocturnal energisation of the tonoplast to facilitate the inward rectifying uptake of malate (e.g. *AsV-ATPase* and/or *AsV-PPase*) also generally showed CAM induction, with relatively higher transcript levels in leaf tip (Figure 7.1). There is an on-going study in *K. fedtschenkoi* leaves using a transgenic approach, which is aiming to determine whether or not ALMT is the protein responsible for malate transport into the vacuole in the dark (Davies, Boxall, Dever, Knerova and Hartwell, unpublished results). Results to date do support a role for ALMT in the transport of malate in the dark. In addition, the tonoplast dicarboxylate transporter (tDT) has generated unclear results in *K. fedtschenkoi* transgenic lines thus far. Here, in *A. sisalana*, the tDT transcript abundance did not show CAM regulation, but exhibited noticeably higher transcript levels in the pale green leaf base dark samples (**Error! Reference source not found.**). Thus, based on the transcript abundance measurements presented here, ALMT is the most likely candidate to be involved in malate transport into the vacuole, rather than tDT (Figure 7.1).

During the light period, the malic acid accumulated during the preceding dark period is transported out of the vacuole into the cytosol, possibly by tDT (Emmerlich *et al.*, 2003; Hurth *et al.*, 2005; Borland *et al.*, 2009). As already mentioned above, the *As_tDT* transcript level reported here did not show a CAM-associated pattern, but it indicated a high level in the leaf base. On the other hand, *AsALMT* exhibited a good CAM-induced pattern, with up-regulation in the leaf tip light samples. Thus, ALMT may also function in the transport of malate out of vacuole during the light period, as well as the dark influx of malate into vacuole (Figure 7.1).

In *A. sisalana,* malate decarboxylation is most likely to be processed via mitochondrial NAD-ME, as *AsNAD-ME* was found among the CAM genes expressing higher transcript levels in the leaf tip, whereas *AsNADP-ME* did not exhibit a CAM-induction, and was detected with lower FPKM values indicating it was present only as a rarer transcript than *AsNAD-ME*. The use of NAD-ME in the light for malate decarboxylation would require the mitochondrial dicarboxylate carrier (*AsDIC*) to transport malate into the mitochondrion in the light (Palmieri *et al.*, 2008). For the plants employing mitochondrial NAD-ME route, this would also require the putative mitochondrial pyruvate transporter (*AsMPC*) in order for the pyruvate resulting from NAD-ME activity to be transported out of the mitochondrion (Bricker *et al.*, 2012) into the cytosol where PPDK (PPDK may also be in the chloroplast) can use the pyruvate to generate PEP which can in turn enter gluconeogenesis. Generally, transcript levels of enzymes involved in light malate decarboxylation and transport showed a clear CAM pattern, particularly *AsDIC,* which exhibited a considerably higher level of transcript in the leaf tip samples. In addition, there is also a dicarboxylate/ tricarboxylate carrier (DTC) that could also transport malate into the mitochondrion in exchange for citrate, isocitrate or aconitate from the TCA cycle (Picault *et al.*, 2002). However, *AsDTC* did not show a CAM-induced pattern in *A. sisalana* (data not shown).

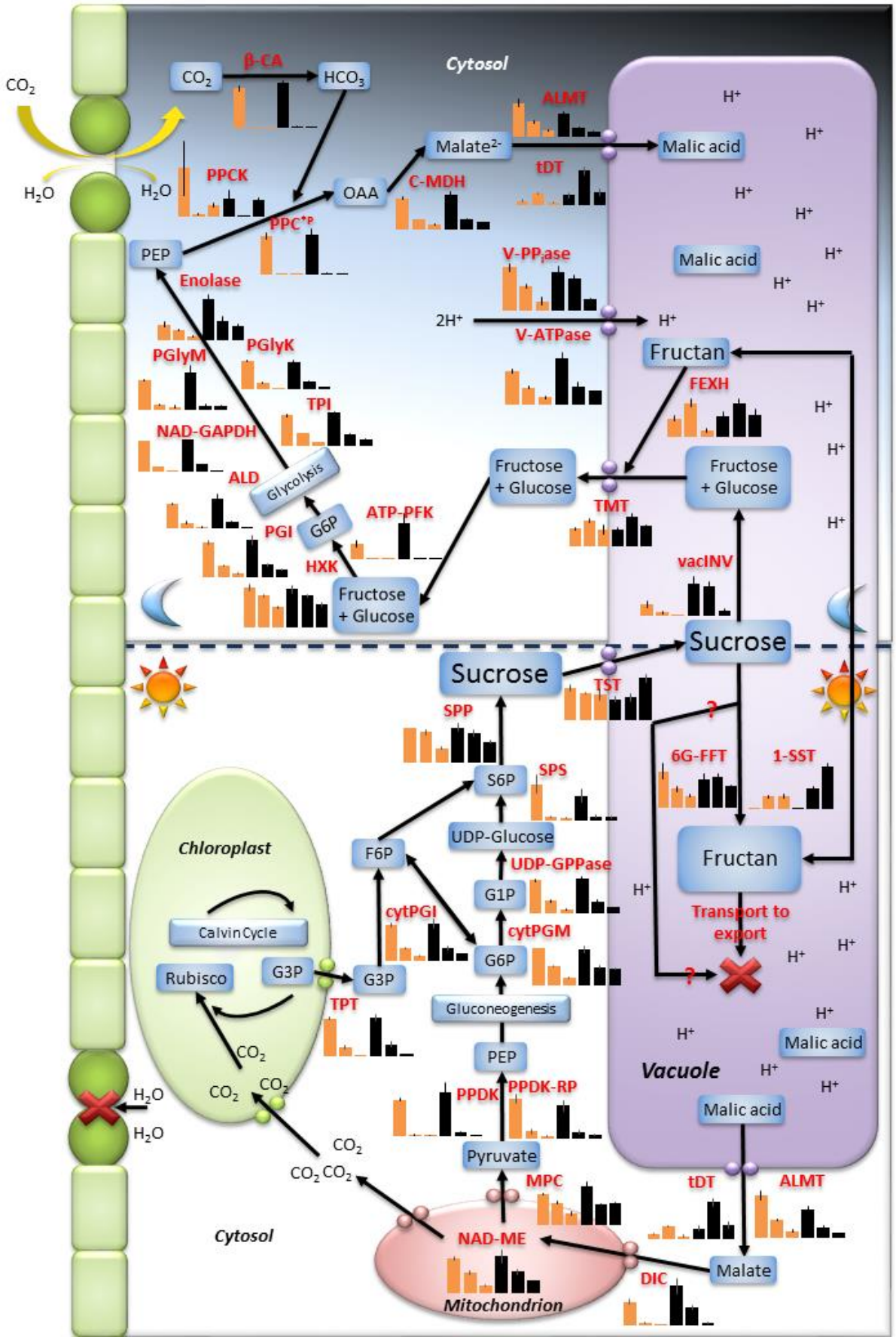

Cytosol
β-CA
CO2
HCO3
ALMT
Malate2-
Malic acid
tDT
PPCK
C-MDH
OAA
PPC*P
PEP
V-PPiase
Malic acid
Enolase
2H+
V-ATPase
Fructan
PGlyM
PGlyK
FEXH
TPI
NAD-GAPDH
Fructose + Glucose
Fructose + Glucose
ALD
Glycolysis
TMT
ATP-PFK
PGI
G6P
vacINV
HXK
Fructose + Glucose
Sucrose
Sucrose
TST
SPP
6G-FFT
1-SST
S6P
SPS
Chloroplast
UDP-Glucose
F6P
Fructan
UDP-GPPase
G1P
Transport to export
cytPGI
cytPGM
Calvin Cycle
G6P
Rubisco
G3P
G3P
Gluconeogenesis
TPT
Malic acid
PEP
Vacuole
PPDK
PPDK-RP
Malic acid
Pyruvate
tDT
ALMT
MPC
NAD-ME
Malate
DIC
Cytosol
Mitochondrion
H2O
CO2
H+

Figure 7.1 *In silico* prediction of the CAM pathway in the mature leaf tip in *A. sisalana* based on RNA-seq derived transcript abundance patterns for the associated genes.

A newly proposed diagram of the CAM pathway in *A. sisalana* heavily modified from the pathway presented by Borland *et al.*, (2009). The enzymes that catalyse the reactions are highlighted in red. Each enzyme is coupled below with transcript abundance (FPKM) bar charts of its encoding gene. Orange bar charts represent light samples including leaf tip, base and white segment from left to right. Black bar charts represent dark samples including leaf tip, base and white segment from left to right. Black arrows represent reactions. The dashed line running across the centre divides the reactions into dark at the top and light at the bottom. The thick green attached blocks to the left of the diagram represent the leaf epidermis. The gap between green circles represents a stomatal pore.

The decarboxylation of malate in the light yields CO_2 and pyruvate or PEP (depending on the species and major malate decarboxylation enzymes). CO_2 is re-fixed via Rubisco in the Calvin-Benson cycle and pyruvate is converted to PEP by pyruvate, orthophosphate dikinase (PPDK) (Evans and Wood, 1968). PPDK is phosphorylated or dephosphorylated by PPDK regulatory protein (PPDK-RP) resulting in inactivation and activation of PPDK respectively (Astley *et al.*, 2011). PEP together with G3P, a product from Calvin-Benson cycle efflux-transported by the triose phosphate/phosphate translocator (TPT) (Walters *et al.*, 2004), is metabolised through gluconeogenesis to produce sugars, which in the case of *A. sisalana* are proposed here to be mainly sucrose. As presented in **Error! Reference source not found.**, cytosolic PGI catalyses the conversion of glucose 6-phosphate (G6P) into fructose 6-phosphate (F6P) (Topper, 1957), and cytosolic PGM converts G6P into glucose 1-phosphate (G1P) (Ray and Roscelli, 1964). UDP-glucose pyrophosphorylase (UDP-GPPase) catalyses the conversion of G1P into UDP-glucose. UDP-glucose is then the substrate for sucrose phosphate synthase (SPS) which converts UDP-glucose and D-fructose 6-phosphate (F6P) into sucrose-6-phosphate (S6P) for sucrose synthesis (Mendicino, 1960) by the activity of sucrose phosphate phosphatase (SPP). Overall, the transcript abundance of the genes encoding enzymes involved in gluconeogenesis and sucrose synthesis showed strong CAM induction, including clear results for *AsPPDK*, *AsPPDK-RP*, *AsTPT*, *AsUDP-GPPase* and *AsSPS* (Figure 7.1). *AsPPDK* showed the strongest CAM induction, with a significantly higher level of transcript in leaf tip relative to the other parts of

the leaf (Figure 7.1). The rest of the enzymes, including cytosolic *AsPGI* and *AsPGM*, and *AsSPP* also exhibited a good level of CAM induction in terms of their transcript levels (Figure 7.1). *SPP* showed one of the lowest levels of CAM-induction for its transcript level comparing the leaf tip with the other sections, when compared to the other enzymes, indicating that sucrose synthesis also occurs in other parts of the leaf. In terms of light/ dark regulation, there were relatively small changes in transcript abundance between the light and dark samples for most of the "decarboxylation pathway" genes, except for *AsSPS* and *AsPPDK-RP* whose transcript levels were noticeably higher in the light compared to the dark in the leaf tip (**Error! Reference source not found.**). These patterns of light/ dark cycling in transcript abundance were consistent with the fact that these reactions occur during the light period.

Sucrose, after being biosynthesised in the cytosol, is subsequently transported via a tonoplast sugar transporter (*As_TST*) into the vacuole (Jung *et al.*, 2015), where it is stored until required. Based on the results in this study, together with the study from Albaijan and Borland, (unpublished) and Christopher and Holtum, (1996), who studied several diverse CAM species including an Agave species, sucrose is proposed and most likely to be remobilised at night to provide the PEP for the nocturnal CO_2 fixation by PEPC in Agave. The induction of the vacuolar invertase (vacINV) would be expected in the leaf tip if sucrose is used as the carbohydrate source for CAM in Agave, as hydrolysis of sucrose at the beginning of the dark period via vacINV would allow sucrose mobilisation for PEP biosynthesis in the dark (Tauzin *et al.*, 2014). Here, *As_vacINV* exhibited relatively strong CAM induction for its transcript pattern (higher transcript level in leaf tip in the dark samples; **Error! Reference source not found.**). The activity of vacINV yields hexoses such as glucose and fructose. These hexoses are then transported out of vacuole into cytosol possibly by tonoplast monosaccharide transporter (TMT) (Wormit *et al.*, 2006). The role of TMT in transporting vacuolar sugar is supported by the study which proposed such a role in *A. comosus* leaves (McRae *et al.*, 2002). The sugar

transport enzymes presented here, *As_TST* and *As_TMT*, did not show CAM induction in their transcript level (they were not consistently higher in the leaf tip; **Error! Reference source not found.**). They are however likely to be involved in the transport of sugars into and out of the vaculoe throughout the whole leaf of the plant, and thus it may not be likely that these genes would be solely induced in the tip relative to the other assayed leaf tissues.

Glucose and fructose transported from vacuole into cytosol are then mobilised through glycolysis which involves various activities and enzymes including as hexokinase (HXK), ATP-phosphofructokinase (ATP-PFK), PGI, fructose-bisphosphate aldolase (ALD), triose-phosphate isomerase (TPI), NAD-glyceraldehyde 3-phosphate dehydrogenase (NAD-GAPDH), phosphoglycerate kinase (PGlyK), phosphoglycerate mutase (PGlyM), and enolase (Berg *et al.*, 2002). This process yields and provides PEP as substrate for the dark CO_2 fixation. Overall, transcript levels of genes encoding these enzymes showed relatively good levels of CAM induction, including *AsATP-PFK*, *AsALD*, *AsNAD-GAPDH*, and *AsPGlyM*, exhibiting much higher level of transcript in leaf tip relative to the other leaf sections (**Error! Reference source not found.**). These genes also tended to be more highly abundant transcripts for the dark samples compared to the light samples (**Error! Reference source not found.**), which correlates well with their dark activities. However, the difference in terms of leaf development and light/ dark regulation of *AsHXK* transcript was not as pronounced as for the other enzymes. It is likely that *AsHXK* will be involved in other activities in other parts of the leaf.

The metabolic analysis of this study and the newly proposed CAM pathway have mainly focused on sucrose (**Error! Reference source not found.**) as, from what has been known, it is the most likely source of carbon for PEP provision in the dark for PEPC catalysed CO_2 uptake (Wang and Nobel, 1998). However, fructans were also one possible option for the source of carbohydrate for PEP provision in the dark for CAM. As there was no previous data

characterising the source of carbohydrate used for CAM in *A. sisalana*, it is important to focus on the regulation of genes related to sucrose synthesis and breakdown, and the regulation of the genes related to fructan synthesis and breakdown within the RNA-seq dataset generated here, in order to improve understanding of the most likely pathway that could happen in the dark during carbohydrate breakdown. Thus, the fructan-related genes were identified and added into the newly proposed pathway (**Error! Reference source not found.**).

Unpublished work by Albaijan and Borland, (unpublished) revealed that some CAM Agave species showed a decline in fructan content in the dark against a background of a large level of nocturnal sucrose turnover. If Agave stores and uses fructans for nocturnal PEP supply for nocturnal CO_2 fixation in addition to sucrose, *AsFEXH* might play an important role for breaking down of fructans into fructose (Krivorotova and Sereikaite, 2014) in order to provide substrate for glycolysis that yields PEP. However, the transcript level of *AsFEXH* presented here does not seem to show CAM induction. It is higher in leaf base than other segments (**Error! Reference source not found.**). This might be due to the possibility that fructan is degraded in leaf base and/ or stored there and temporally turned over there to fuel respiration and growth. The transcript abundance of *AsFEXH* in the leaf tip in the dark might reflect the activity of fructan degradation to supply PEP, in addition to the activity of vacINV in degrading sucrose, which is proposed to be main carbohydrate source for PEP supply. In *A. deserti*, fructan synthesis was found to occur in the vascular tissue of mature leaves (Wang and Nobel, 1998). It is believed that fructans are generally synthesized in the vacuole through the activity of the enzyme fructosyl transferase utilising sucrose that is imported as substrate (Valluru and Van den Ende, 2008). The examples of well-defined fructan synthetic genes include fructan:fructan 1-fructosyltransferase (1-FFT), sucrose:fructan 6-fructosyltransferase (6SFT) and sucrose:sucrose 1-fructosyltransferase (1-SST) (Avila de Dios *et al.*, 2015). In this study, *As1-SST* showed a transcript level that was higher in the leaf base and white segment

relative to the leaf tip (**Error! Reference source not found.**). It is possible that fructan is mainly synthesised in leaf base and white section, using sucrose that is most likely to have been transported from the photosynthetic source tissues at the leaf tip. The synthesised fructans could then be accumulated and stored in the leaf base and white section, which are likely to be sink tissues that receive sucrose from photosynthesis in the tip. In wheat (*Triticum aestivum*), 1-SST is important in fructan accumulation in leaf tissues during cold periods (Kawakami and Yoshida, 2002). 6G-FFT is also found to have an activity for fructan synthesis (Kawakami and Yoshida, 2002). Unlike *As1-SST*, *As6G-FFT* transcripts showed a reasonable level of CAM induction (Figure 7.1). This indicates that fructans are also possibly synthesised in the CAM-performing leaf tip and degraded via FEXH for the carbohydrate supply of nocturnal PEP synthesis, and/ or transported out of the leaf tip into the leaf base or white section for long-term storage. However, mechanisms behind the process of fructan transport into sink tissues and accumulation in the vacuole are still undiscovered.

To investigate whether sucrose turnover was sufficent to supply the nocturnal PEP required for CO_2 fixation to malate, the nocturnal decrease in sucrose was calculated in moles and converted to moles of PEP produced by multiplying by 4 in order to obtain the potential malate production that the measured sucrose could provision. One sucrose (12-carbon) can be broken down through glycolysis to four PEP (3-carbon), which can be then utilised in the dark CO_2 fixation to produce four malates (4-carbon) (Wang and Nobel, 1998). The amount of sucrose turnover during the dark period was 11.92 µmole of sucrose. If all of this was used for PEP production and subsequent CO_2 fixation to malate, then the sucrose turnover could support 47.68 µmole of malate production by dawn. Thus, the turnover of sucrose alone was clearly not sufficient to support the amount of malate accumulation that was measured (172.57 µmole) due to dark CO_2 fixation (difference between 22h and 14h; Table 7.1). These findings support the conclusion that in *A. sisalana* sucrose turnover alone was insufficient to

serve as the nocturnal source of carbohydrates required to supply substrate for CAM CO_2 fixation in the dark. Therefore, fructan (described in the previous paragraph; Figure 7.1) and/ or other forms of carbohydrates such as starch may also be used to generate PEP supply for nocturnal CO_2 fixation. This will be an important area for future research as the source of carbohydrates for provisioning nocturnal PEP and CO_2 fixation is a key area of CAM research where many questions still remain to be answered, especially in terms of the circadian control of carbohydrate turnover and PEP supply.

Table 7.1 Theoretical maximal nocturnal malate production calculated from the nocturnal decrease in leaf sucrose and compared to the measured nocturnal malate production in the youngest fully expanded *A. sisalana* leaf tip.

	Tip µmole/g fresh weight		
Time	**Malate**	**Sucrose**	**Sucrose*4 = malate**
2h	259.43	9.55	38.21
6h	79.26	21.38	85.53
10h	88.18	20.94	83.78
14h	131.67	24.27	97.10
18h	204.24	21.02	84.09
22h	304.24	12.35	49.41
Nocturnal decrease of sucrose for malate production	-	24.27-12.35 = 11.92	**47.68**
Nocturnal malate production	304.24-131.67 = **172.57**	-	-

In addition, Avila de Dios *et al.*, (2015) failed to identify through *in silico* analysis of *de novo* transcriptome data for several *Agave* species any assembled cDNA sequences that corresponded to 1-FFT and 6-SFT. They studied four different Agave species (*A. tequilana, A. deserti, A. victoriae-reginae,* and *A. striata*), suggesting that these enzymes may be lacking across a wide range of members of the genus. They argued that those enzymes should be required to produce the types of fructans found in Agave, and yet they were not able to find sequences that could be identified through sequence homology as having the 6-SFT and 1FFT activity. In this study, the alignment search was performed against *A. sisalana* Illumina

transcriptome using well-defined protein sequences for 6-SFT from wheat (*Triticum aestivum*) (Diedhiou *et al.*, 2012) and 1-FFT from *Arctium lappa* (Ueno *et al.*, 2011). The search results did not return any good candidate contigs for both proteins, being consistent with the finding of Avila de Dios *et al.*, (2015). They further explained that it was possible that, in Agave species, the expression levels of genes encoding 1-FFT are very low or tissue-specific, with the limitation being that the tissues containing transcripts for these genes might not have been sampled. The difficulty to identify good candidate contigs encoding 6-SFT might also be due to the same possible explanation as proposed for 1-FFT.

Furthermore, there is a need to be careful with the identification of the FEXH gene(s) as they are closely related to the invertases (Le Roy *et al.*, 2007), which belong to the same large protein family, glycoside hydrolase family 32 (PGHF32). In *A. thaliana*, the identification of the enzymes of this protein family based on sequence data only has led to mis-annotation of *AtCwinv3,* which was later revealed to be involved in FEXH activity (De Coninck *et al.*, 2005). Thus, multiple-alignment of protein sequences was performed and a phylogenetic tree was generated using Geneious (Drummond *et al.*, 2011) (Figure 7.2). The tree is a neighbour-joining tree generated with bootstrap settings set to 100 repetitions in order to distinguish between the identified contigs of FEXH and invertase, and to find the best possible putative candidate contig(s) for those genes from the *A. sisalana* transcriptome assembled and reported here. The following protein sequences from *A. sisalana* genes: *AsFEXH (c538578_g1_i1*), *invertase (AsINV*; *c567445_g2_i1*), *vacuolar invertase* (*As_vacINV; c571529_g1_i2*), and *cell wall invertase* (*As_cwINV*; *c589904_g1_i1*) were identified using the reciprocal best BLAST hit methods described above and were then aligned with the well-defined, published protein sequences for *FEXH* and *INV* genes from other species including *AbFEXH* (AT5G11920.1; Quilliam *et al.*, (2006)) and *Ab*_cwINV (AT1G12240.1; Carter *et al.*, (2004)) from *A. thaliana*, *LpFEXH* (AAZ29514.1; Chalmers *et al.*, (2005)) from *Lolium perenne,*

and *Gh_vacINV* (ACQ82802.1; Wang *et al.*, (2010)) from *Gossypium hirsutum*. Also, recently identified sequences encoding *AtFEXH* (KR138454.1), *At_cwINV* and *AtINV* (KR138451.1; KR138450.1) from *A. tequilana,* and *Av_vacINV* (KR138448.1) from *A. victoriae-reginae* were included in the multiple-protein sequence alignments used for the phylogenetic tree analysis (Avila de Dios *et al.*, 2015). The results of the multiple alignment and phylogenetic tree demonstrates that *A. sisalana AsFEXH* gene (contig c538578_g1_i1) showed the closest relationship to the *FEXH*s from other species, confirming that the identified putative *A. sisalana AsFEXH* gene presented here is not closely related to other known invertase genes. Thus, it should be a good candidate for *A. sisalana FEXH* (Figure 7.2).

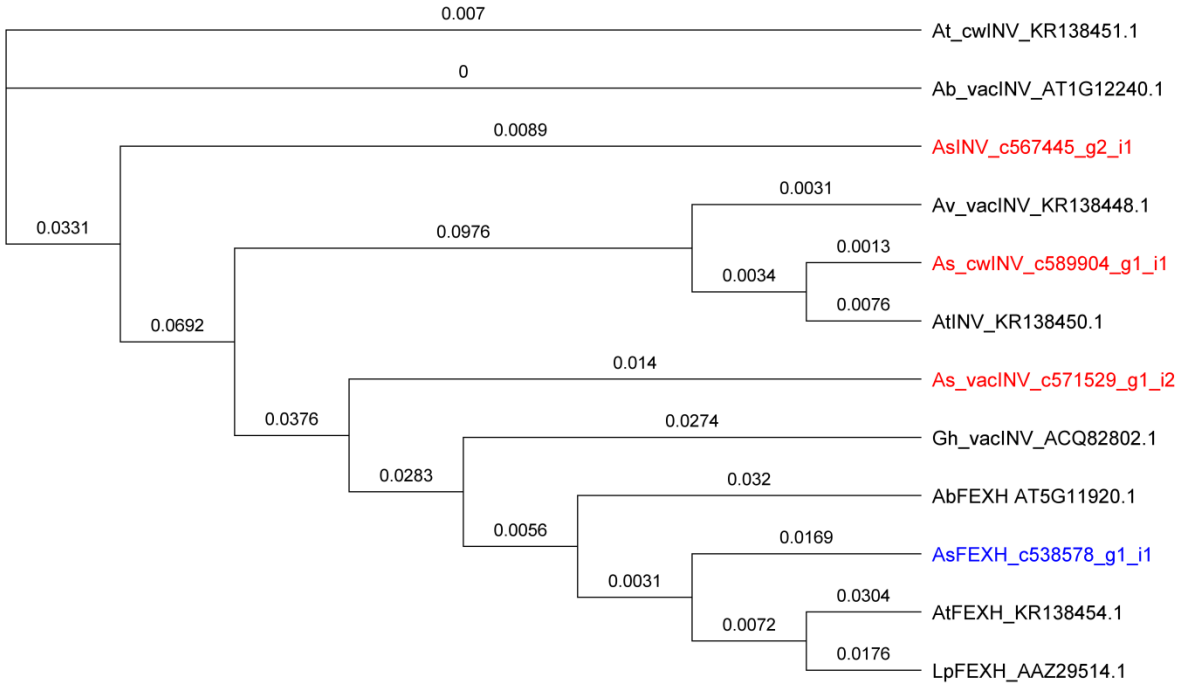


Figure 7.2 Phylogenetic tree of *FEXH* and *INV* genes in different species

The multiple alignment of protein sequences was performed and phylogenetic tree was generated using Geneious (Drummond *et al.*, 2011) with bootstrap setting of 100 repetitions. Red and blue gene/ contig codes were derived from the Illumina transcriptome of *A. sisalana* generated in this study. The rest of the amino acid sequences were derived from online database: GeneBank (Benson *et al.*, (2013); http://blast.ncbi-.nlm.nih.gov/Blast.cgi) and Arabidopsis (Lamesch *et al.*, (2011); https://www.arabidopsis.org/Blast/). Numbers indicate substitutions per site. At = *A. tequilana*, As = *A. sisalana*, Av = *A. victoriae-reginae*, Ab = *Arabid-opsis thaliana*, Gh = *Gossypium hirsutum* and Lp = *Lolium perenne*.

Predicted sub-cellular localisation of key steps in the CAM pathway in *A. sisalana*

A further interesting experiment was to determine the *in silico* predicted localisation of *A. sisalana* NAD-ME, NADP-ME, NAD-MDH and PPDK identified in this study as candidate CAM genes. The protein sequences obtained from the longest ORF of the candidate contigs of the genes encoding these enzymes were analysed using the following localisation prediction tools: Plant-mPLoc, (Chou and Shen, 2010), ChloroP (Emanuelsson *et al.*, 1999), TargetP (Emanuelsson *et al.*, 2000), WoLF PSORT (Horton *et al.*, 2007), and Predotar (Small *et al.*, 2004) (Table 7.2). The results for the predicted localisation of *AsNAD-ME* indicated clearly, with consistent results among the various tools, that it is most likely to be localised to the mitochondrion. This result is consistent with the study in CAM plant, *K. daigremontiana*, which reported that NAD-ME is exclusively present in mitochondria and not in chloroplasts or the cytoplasm (Dittrich, 1976). Furthermore, mitochondrial NAD-ME has also been demonstrated recently to be the major malate decarboxylation enzyme in the CAM leaves of *K. fedtschenkoi* (Dever *et al*., 2015).

AsNAD-MDH, however, showed inconsistent results across the different prediction tools. The protein sequence of the longest ORF from the CAM-induced contig (Figure 7.1) aligned well with the *AsNAD-MDH* full length CDS protein sequence from *A. thaliana*, confirming that the *A. sisalana* protein sequence used is full length with the presence of N-terminal elements (at the start of the protein sequence) where the prediction tools target for the analysis. In this study, *NAD-MDH* is proposed to be localised in the cytosol.

The protein sequences for *AsNADP-ME* and *AsPPDK*, including all of the isoforms in the current assembly, lacked the N-terminal regions when aligned with the full length CDS sequences from *A. thaliana*. This might have an effect on the results of the prediction tools as they rely on N-

terminal targeting sequences (Table 7.2). PPDK has been proposed to be a chloroplastic enzyme in the inducible-CAM species *M. crystallinum* (Kondo *et al.*, 1998). It is clear that further work will be required in the future in order to decipher the exact sub-cellular localisation of several of the key CAM enzymes identified here, including cloning of the complete full length cDNAs encoding each CAM-associated member of these genes families, and biochemical work using isolated mitochondria, chloroplast and cytosolic extracts in order to determine which compartment(s) contain the highest activities of the various CAM enzymes.

Table 7.2 Subcellular localisation predictions for *AsNAD-ME*, *AsNADP-ME*, *AsNAD-MDH* and *AsPPDK* using various prediction tools.

		Tools (results and scores)				
Gene	**Contig**	**Plant-mPLOC**	**ChloroP**	**TargetP**	**WOLF PSORT**	**Predotar**
AsNAD-ME	c453533_g1_i1	Mitochondrion	No 0.491	M 0.914	chlo: 8, mito: 6	Mitochondrial 0.76
AsNADP-ME	c558461_g1_i1	Chloroplast	No 0.440	Other 0.520	chlo: 10, nucl: 2, plas: 1	Elsewhere 0.71
AsNAD-MDH	c565688_g2_i1	Chloroplast Mitochondrion	No 0.435	Other 0.573	cyto: 8, chlo: 3.5, chlo_mito: 2.5, nucl: 1	Elsewhere 0.97
AsPPDK	c525272_g3_i1	Chloroplast	No 0.486	Other 0.676	chlo: 6, cyto: 3, mito: 2, pero: 2	Elsewhere 0.97

In conclusion, the *de novo* transcriptome analysis of *A. sisalana* generated in this study illustrates the effectiveness of short-read, next-generation RNA-sequencing technology for the rapid identification of genes involved in the a biological process of interest (Martin and Wang, 2011). This work generates a high-quality resource for Agave transcriptome profiling. The comprehensive RNA-seq data analysis, including the differential expression (DE) method used here, facilitates the detailed and in-depth exploration of the regulation of the transcriptome both in terms of leaf development and light/ dark regulation. This permitted the discovery of several novel CAM-induced and circadian clock-controlled transcription factor gene

candidates, which will aid future studies of CAM regulation in *A. sisalana*. Moreover, the newly proposed CAM pathway (Figure 7.1) provides a robust and testable model for the CAM pathway in *A. sisalana* and its regulation over the light/ dark cycle. It also provides insights into the carbohydrate source used for PEP provision in the dark for CAM, especially in relation to the possible subsidiary role of fructan metabolism. The proposed Agave CAM pathway is relatively simple relative to the pathway proposed for other CAM models (J. Hartwell, personal communication). Thus, Agave seems to be an easier model for the engineering of a CAM system into C_3 plants compared to other model CAM species such as *K. fedtschenkoi*. Considering their potential as a sustainable source of renewable biomass that can be used for biofuel/ bioenergy, Agave would also suit better when considering engineering CAM into C_4 due to the fact that plants employing CAM and C_4 have high degree of similarity in mechanism and productivity. Thus, *A. sisalana* could be a good CAM biofuel crop model.

Future work

To gain further insights into the mechanism of CAM in Agave and its daily regulation in response to the light/ dark cycle and the endogenous circadian clock, the novel transcription factors genes identified from the RNA-seq data analysis could be processed further into manipulating the best possible novel CAM gene(s) by employing transgenic approaches. The function of the selected genes could be tested using RNAi knock-down loss-of-function techniques in order to determine *in planta* function of each gene (Dever et al., 2015). Such study would be the first of its kind in any Agave species. It would also be beneficial to develop tissue culture techniques for the efficient regeneration of *A. sisalana* through callus induction and subsequent root and shoot induction; attempts to establish various explants of *A. sisalana* in tissues culture and to induce callus formation were not successful as part of a preceding masters project that was undertaken in the year before this PhD project started (Bupphada

and Hartwell, unpublished). An efficient tissue culture based regeneration system would be a key prerequisite for any future work with transgenic approaches in *A. sisalana*. Due to the time limits of this PhD project (4 years with 3.5 years of bench work), the number of novel CAM-induced transcription factors that could be identified from the RNA-seq analysis and studied in further detail for possible circadian clock control, had to be limited. Thus, it would be worthy to identify more candidate transcription factor genes and study the regulation of their transcript abundance in more detail through Q-RT-PCR in order to screen for best candidate for the further study such as the transgenic approaches mentioned above.

Alongside the plant molecular biology approaches, other future work is to screen microbial fermentative species for their capacity to grow at high cell density on different components of Agave biomass with concomitant bioconversion by fermentation of soluble sugars, fructans or cellulose to biofuels like ethanol or long chain alkanes. Other future work could also include a study of fructan metabolism related genes and, if possible, *A. sisalana* end product processing such as bioethanol production. Moreover, studies using $^{13}CO_2$ or $H^{13}CO3^-$ would also be an interesting area for further study in the future. Pulse-chase experiments in which the *A. sisalana* leaf tip was allowed to fix the ^{13}C form in the dark would allow the metabolite measurements, for example using ^{13}C NMR measurements, that would in turn allow the pathway of daily carbon flow during CAM in *A. sisalana* to be investigated in much greater detail than the current level of understanding. Such studies would also allow the tracking of sugars as they are moved around the leaf from the photosynthetic source in the leaf tip to the sink tissues where growth occurs in the leaf base.

It would also be interesting in the future to investigate CAM in different cell types of the Agave leaf in addition to the leaf developmental profile studied here, as it has been shown in this study that some known-CAM genes such as *AsPPCK* failed to exhibit a clear and constant

pattern of leaf developmental control between the base and the tip. Such genes might be more clearly differentially expressed in subsets of cells; for example if the green photosynthetic chlorenchyma cells could be sampled and compared and contrasted with the underlying white, non-photosynthetic water-storage tissues that occupies most of the internal volume of the highly succulent *A. sisalana* leaf. Additionally, it would also be interesting to investigate the gradient development along the whole Agave leaf from white basal part to tip to see all the developmental stages from C_3 to CAM (i.e. by using samples collected every cm from base to tip). Lastly, the *de novo* transcriptome assembly and quantitative data generated in this study is useful for developing molecular markers for further efforts in Agave breeding or other molecular studies via the identification of novel polymorphic sites such as SNPs and indels. Such molecular markers could be linked with key productivity and water use efficiency traits using techniques such as marker assisted selection, and thus improved varieties of Agave could be generated more rapidly by tracking the key molecular markers that associate with important agronomic traits. However, the slow life cycle of Agave, often 10-12 year's growth before flowering, and lack of seed set in certain cultivated species of Agave including *A. sisalana*, suggest that transgenic approaches may be a more facile method for the future improvement of species such as *A. sisalana*. Considering that *A. sisalana* is not used as a source of food, it may be that transgenic versions of this crop would be more readily accepted by society, as long as the appropriate controls were put in place to prevent the spread of transgenes to wild relatives.

It has been predicted that global warming driven by human emissions of fossil fuel derived CO_2 will lead to a hotter and often drier world in many regions of the earth where water is already a limiting resource for agriculture. Agaves have great potential as a water use efficient source of biomass for bioenergy and/ or renewable chemicals for industry (Borland *et al.*, 2009; Stewart, 2015; Yang *et al.*, 2015). Further understanding of the genomes, transcriptomes,

proteomes, metabolomes and phenomes of Agaves will greatly facilitate their further development and improvement for delivering sustainable and renewable biomass in regions of the world where the current major food crop species, such as rice and wheat, would struggle to grow productively. Thus, it is hoped that the work described in this thesis provides a solid foundation for the further development of Agave in the future.

References

Adams, P., Nelson, D.E., Yamada, S., Chmara, W., Jensen, R.G., Bohnert, H.J. & Griffiths, H. (1998). Growth and development of Mesembryanthemum crystallinum (Aizoaceae). *New Phytologist,* **138,** 171-190.

Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol,* **11,** R106.

Andersen, C.L., Jensen, J.L. & Orntoft, T.F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research,* **64,** 5245-5250.

Antony, E., Taybi, T., Courbot, M., Mugford, S.T., Smith, J.a.C. & Borland, A.M. (2008). Cloning, localization and expression analysis of vacuolar sugar transporters in the CAM plant Ananas comosus (pineapple). *Journal of Experimental Botany,* **59,** 1895-1908.

Aosaar, J., Varik, M. & Uri, V. (2012). Biomass production potential of grey alder (Alnus incana (L.) Moench.) in Scandinavia and Eastern Europe: A review. *Biomass and Bioenergy,* **45,** 11-26.

Arizaga, S. & Ezcurra, E. (1995). Insurance against reproductive failure in a semelparous plant: bulbil formation in Agave macroacantha flowering stalks. *Oecologia,* **101,** 329-334.

Aronesty, E. (2013). Comparison of sequencing utility programs. *Open Bioinformatics Journal,* **7,** 1-8.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S. & Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics,* **25,** 25-29.

Astley, H.M., Parsley, K., Aubry, S., Chastain, C.J., Burnell, J.N., Webb, M.E. & Hibberd, J.M. (2011). The pyruvate, orthophosphate dikinase regulatory proteins of Arabidopsis are both bifunctional and interact with the catalytic and nucleotide-binding domains of pyruvate, orthophosphate dikinase. *Plant Journal,* **68,** 1070-1080.

Avila De Dios, E., Gomez Vargas, A.D., Damián Santos, M.L. & Simpson, J. (2015). New insights into plant glycoside hydrolase family 32 in Agave species. *Frontiers in Plant Science,* **6,** 594.

Bartholomew, D.M., Rees, D.J., Rambaut, A. & Smith, J.A. (1996). Isolation and sequence analysis of a cDNA encoding the c subunit of a vacuolar-type H(+)-ATPase from the CAM plant Kalanchoe daigremontiana. *Plant Mol Biol,* **31,** 435-442.

Bennet-Clark, T. (1933). The Rôle of the Organic Acids in Plant Metabolism Part I. *New Phytologist,* **32,** 37-71.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2013). GenBank. *Nucleic Acids Research,* **41,** D36-42.

Berg, J.M., Tymoczko, J.L. & Stryer, L. (2002). Glycolysis is an energy-conversion pathway in many organisms. In: *Biochemistry*. W H Freeman, New York.

Black, C. & Osmond, C.B. (2003). Crassulacean acid metabolism photosynthesis: `working the night shift'. *Photosynthesis Research,* **76,** 329-341.

Black, C.C., Chen, J.Q., Doong, R.L., Angelov, M.N. & Sung, S.J.S. (1996). Alternative Carbohydrate Reserves Used in the Daily Cycle of Crassulacean Acid Metabolism. In: *Crassulacean Acid Metabolism* (ed. by K. Winter & J.A. Smith), pp. 31-45. Springer Berlin Heidelberg.

Borland, A.M., Griffiths, H., Hartwell, J. & Smith, J.A. (2009). Exploiting the potential of plants with crassulacean acid metabolism for bioenergy production on marginal lands. *Journal of Experimental Botany,* **60,** 2879-2896.

Borland, A.M., Hartwell, J., Jenkins, G.I., Wilkins, M.B. & Nimmo, H.G. (1999). Metabolite Control Overrides Circadian Regulation of Phosphoenolpyruvate Carboxylase Kinase and CO2 Fixation in Crassulacean Acid Metabolism. *Plant Physiology,* **121,** 889-896.

Borland, A.M., Hartwell, J., Weston, D.J., Schlauch, K.A., Tschaplinski, T.J., Tuskan, G.A., Yang, X. & Cushman, J.C. (2014). Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends in Plant Science,* **19,** 327-338.

Borland, A.M. & Taybi, T. (2004). Synchronization of metabolic processes in plants with Crassulacean acid metabolism. *Journal of Experimental Botany,* **55,** 1255-1265.

Borland, A.M., Wullschleger, S.D., Weston, D.J., Hartwell, J., Tuskan, G.A., Yang, X. & Cushman, J.C. (2015). Climate-resilient agroforestry: physiological responses to climate change and engineering of crassulacean acid metabolism (CAM) as a mitigation strategy. *Plant Cell Environ*. **38(9)**,1833-49

Bousios, A., Saldana-Oyarzabal, I., Valenzuela-Zapata, A.G., Wood, C. & Pearce, S.R. (2007). Isolation and characterization of Ty1-copia retrotransposon sequences in the blue agave (Agave tequilana Weber var. azul) and their development as SSAP markers for phylogenetic analysis. *Plant Science,* **172,** 291-298.

Bowers, J.E., Webb, R.H. & Rondeau, R.J. (1995). Longevity, recruitment and mortality of desert plants in Grand Canyon, Arizona, USA. *Journal of Vegetation Science,* **6,** 551-564.

Boxall, S.F., Foster, J.M., Bohnert, H.J., Cushman, J.C., Nimmo, H.G. & Hartwell, J. (2005). Conservation and divergence of circadian clock operation in a stress-inducible Crassulacean acid metabolism species reveals clock compensation against stress. *Plant Physiology,* **137,** 969-982.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., Mccurdy, S., Foy, M. & Ewan, M. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology,* **18,** 630-634.

Bricker, D.K., Taylor, E.B., Schell, J.C., Orsak, T., Boutron, A., Chen, Y.-C., Cox, J.E., Cardon, C.M., Van Vranken, J.G., Dephoure, N., Redin, C., Boudina, S., Gygi, S.P., Brivet, M.,

Thummel, C.S. & Rutter, J. (2012). A Mitochondrial Pyruvate Carrier Required for Pyruvate Uptake in Yeast, Drosophila, and Humans. *Science,* **337,** 96-100.

Broetto, F., Ttge, U. & Ratajczak, R. (2002). Influence of light intensity and salt-treatment on mode of photosynthesis and enzymes of the antioxidative response system of Mesembryanthemum crystallinum. *Functional Plant Biology,* **29,** 13-23.

Bryant, S. & Manning, D.L. (1998). Formaldehyde Gel Electrophoresis of Total RNA. In: *RNA Isolation and Characterization Protocols* (ed. by R. Rapley & D.L. Manning), pp. 69-72 Humana Press, Totowa, New Jersey.

Buchanan-Bollig, I.C. & Smith, J.A. (1984). Circadian rhythms in crassulacean acid metabolism: phase relationships between gas exchange, leaf water relations and malate metabolism in Kalanchoe daigremontiana. *Planta,* **161,** 314-319.

Bustin, S.A. (2010). Why the need for qPCR publication guidelines?--The case for MIQE. *Methods,* **50,** 217-226.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J. & Wittwer, C.T. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry,* **55,** 611-622.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G. & Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal,* **6,** 1621-1624.

Carter, C., Pan, S., Zouhar, J., Avila, E.L., Girke, T. & Raikhel, N.V. (2004). The Vegetative Vacuole Proteome of Arabidopsis thaliana Reveals Predicted and Unexpected Proteins. *The Plant Cell,* **16,** 3285-3303.

Carter, P.J., Nimmo, H., Fewson, C. & Wilkins, M. (1991). Circadian rhythms in the activity of a plant protein kinase. *The EMBO Journal,* **10,** 2063.

Carvalho, R., Feijão, C. & Duque, P. (2013). On the physiological significance of alternative splicing events in higher plants. *Protoplasma,* **250,** 639-650.

Chalmers, J., Lidgett, A., Cummings, N., Cao, Y., Forster, J. & Spangenberg, G. (2005). Molecular genetics of fructan metabolism in perennial ryegrass. *Plant Biotechnology Journal,* **3,** 459-474.

Chase, M.W., Reveal, J.L. & Fay, M.F. (2009). A subfamilial classification for the expanded asparagalean families Amaryllidaceae, Asparagaceae and Xanthorrhoeaceae. *Botanical Journal of the Linnean Society,* **161,** 132-136.

Chastain, C.J., Botschner, M., Harrington, G.E., Thompson, B.J., Mills, S.E., Sarath, G. & Chollet, R. (2000). Further analysis of maize C(4) pyruvate,orthophosphate dikinase phosphorylation by its bifunctional regulatory protein using selective substitutions of the regulatory Thr-456 and catalytic His-458 residues. *Arch Biochem Biophys,* **375,** 165-170.

Chastain, C.J., Fries, J.P., Vogel, J.A., Randklev, C.L., Vossen, A.P., Dittmer, S.K., Watkins, E.E., Fiedler, L.J., Wacker, S.A., Meinhover, K.C., Sarath, G. & Chollet, R. (2002). Pyruvate,Orthophosphate Dikinase in Leaves and Chloroplasts of C(3) Plants Undergoes Light-/Dark-Induced Reversible Phosphorylation. *Plant Physiology,* **128,** 1368-1378.

Chen, L.S., Lin, Q. & Nose, A. (2002). A comparative study on diurnal changes in metabolite levels in the leaves of three crassulacean acid metabolism (CAM) species, Ananas comosus, Kalanchoe daigremontiana and K. pinnata. *Journal of Experimental Botany,* **53,** 341-350.

Chou, K.-C. & Shen, H.-B. (2010). Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE,* **5,** e11335.

Christopher, J.T. & Holtum, J.a.M. (1996). Patterns of Carbon Partitioning in Leaves of Crassulacean Acid Metabolism Species during Deacidification. *Plant Physiology,* **112,** 393-399.

Chu, Y. & Corey, D.R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics,* **22,** 271-274.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S. & Bethel, G. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods,* **5,** 613-619.

Cockburn, W. (1985). Tansley Review No 1. Variation in Photosynthetic Acid Metabolism in Vascular Plants: Cam and Related Phenomena. *New Phytologist,* **101,** 3-24.

Cockburn, W., Ting, I.P. & Sternberg, L.O. (1979). Relationships between Stomatal Behavior and Internal Carbon Dioxide Concentration in Crassulacean Acid Metabolism Plants. *Plant Physiology,* **63,** 1029-1032.

Cook, B.I., Smerdon, J.E., Seager, R. & Coats, S. (2014). Global warming and 21st century drying. *Climate Dynamics,* **43,** 2607-2627.

Corbin, K.R., Byrt, C.S., Bauer, S., Debolt, S., Chambers, D., Holtum, J.a.M., Karem, G., Henderson, M., Lahnstein, J., Beahan, C.T., Bacic, A., Fincher, G.B., Betts, N.S. & Burton, R.A. (2015). Prospecting for Energy-Rich Renewable Raw Materials: Agave Leaf Case Study. *PLoS ONE,* **10,** e0135382.

Covington, M.F., Maloof, J.N., Straume, M., Kay, S.A. & Harmer, S.L. (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biology,* **9**.

Crayn, D.M., Winter, K. & Smith, J.a.C. (2004). Multiple origins of crassulacean acid metabolism and the epiphytic habit in the Neotropical family Bromeliaceae. *Proceedings of the National Academy of Sciences of the United States of America,* **101,** 3703-3708.

Cushman, J.C. & Bohnert, H.J. (1997). Molecular Genetics of Crassulacean Acid Metabolism. *Plant Physiology,* **113,** 667-676.

Cushman, J.C., Tillett, R.L., Wood, J.A., Branco, J.M. & Schlauch, K.A. (2008). Large-scale mRNA expression profiling in the common ice plant, Mesembryanthemum crystallinum, performing C3 photosynthesis and Crassulacean acid metabolism (CAM). *Journal of Experimental Botany,* **59,** 1875-1894.

Dai, A. (2013). Increasing drought under global warming in observations and models. *Nature Climate Change,* **3,** 52-58.

Dalchau, N., Baek, S.J., Briggs, H.M., Robertson, F.C., Dodd, A.N., Gardner, M.J., Stancombe, M.A., Haydon, M.J., Stan, G.B., Gonçalves, J.M. & Webb, A.a.R. (2011). The circadian oscillator gene GIGANTEA mediates a long-term response of the Arabidopsis thaliana circadian clock to sucrose. *Proceedings of the National Academy of Sciences of the United States of America,* **108,** 5104-5109.

Davis, S.C., Dohleman, F.G. & Long, S.P. (2011). The global potential for Agave as a biofuel feedstock. *GCB Bioenergy,* **3,** 68-78.

De Coninck, B., Le Roy, K., Francis, I., Clerens, S., Vergauwen, R., Halliday, A.M., Smith, S.M., Van Laere, A. & Van Den Ende, W. (2005). Arabidopsis AtcwINV3 and 6 are not invertases but are fructan exohydrolases (FEHs) with different substrate specificities. *Plant, Cell & Environment,* **28,** 432-443.

De La Fuente Van Bentem, S. & Hirt, H. (2009). Protein tyrosine phosphorylation in plants: More abundant than expected? *Trends in Plant Science,* **14,** 71-76.

Delgado Sandoval Sdel, C., Abraham Juarez, M.J. & Simpson, J. (2012). Agave tequilana MADS genes show novel expression patterns in meristems, developing bulbils and floral organs. *Sex Plant Reprod,* **25,** 11-26.

Dever, L., Boxall, S., Knerova, J. & Hartwell, J. (2015). Transgenic perturbation of the decarboxylation phase of Crassulacean acid metabolism alters physiology and metabolism but has only a small effect on growth. *Plant Physiology,* **167,** 44-59.

Diedhiou, C., Gaudet, D., Liang, Y., Sun, J., Lu, Z.X., Eudes, F. & Laroche, A. (2012). Carbohydrate profiling in seeds and seedlings of transgenic triticale modified in the expression of sucrose:sucrose-1-fructosyltransferase (1-SST) and sucrose:fructan-6-fructosyltransferase (6-SFT). *J Biosci Bioeng,* **114,** 371-378.

Dittrich, P. (1976). Nicotinamide Adenine Dinucleotide-specific "Malic" Enzyme in Kalanchoe daigremontiana and Other Plants Exhibiting Crassulacean Acid Metabolism. *Plant Physiol,* **57,** 310-314.

Dodd, A.N., Borland, A.M., Haslam, R.P., Griffiths, H. & Maxwell, K. (2002). Crassulacean acid metabolism: plastic, fantastic. *Journal of Experimental Botany,* **53,** 569-580.

Dodd, A.N., Griffiths, H., Taybi, T., Cushman, J.C. & Borland, A.M. (2003). Integrating diel starch metabolism with the circadian and environmental regulation of Crassulacean acid metabolism in Mesembryanthemum crystallinum. *Planta,* **216,** 789-797.

Dodd, A.N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., Hibberd, J.M., Millar, A.J. & Webb, A.a.R. (2005). Plant Circadian Clocks Increase Photosynthesis, Growth, Survival, and Competitive Advantage. *Science,* **309,** 630-633.

Dowson-Day, M.J. & Millar, A.J. (1999). Circadian dysfunction causes aberrant hypocotyl elongation patterns in Arabidopsis. *The Plant Journal,* **17,** 63-71.

Drummond, A.J., Ashton, B., Buxton, S., M., C., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Hava, S., Sturrock, S., Thierer, T. & Wilson, A. (2011). *Geneious v5.4* [Online]. Available: http://www.geneious.com [Accessed 19 June 2012].

Dunlap, J.C., Loros, J.J. & Decoursey, P.J. (2004). *Chronobiology: biological timekeeping,* Sinauer Associates.

Eguiarte, L., Souza, V. & Silva Montellano, A. (2000). Evolución de la familia Agavaceae: filogenia, biología reproductiva y genética de poblaciones. *Boletin de la Sociedad Botanica de Mexico***,** 131-151.

Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol,* **300,** 1005-1016.

Emanuelsson, O., Nielsen, H. & Von Heijne, G. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science,* **8,** 978-984.

Emmerlich, V., Linka, N., Reinhold, T., Hurth, M.A., Traub, M., Martinoia, E. & Neuhaus, H.E. (2003). The plant homolog to the human sodium/dicarboxylic cotransporter is the vacuolar malate carrier. *Proceedings of the National Academy of Sciences of the United States of America,* **100,** 11122-11126.

Engelmann, W. & Johnsson, A. (1998). Rhythms in organ movement. In: *Biological rhythms and photoperiodism in plants* (ed. by P. Lumsden & A. Millar), pp. 35-50. Bios Scientific, Oxford, UK.

Eugenio, S., Bobenrieth, H. & Brian, D.W. (2009). The food price crisis of 2007/2008: Evidence and implications. In: *Joint Meeting of the Intergovernmental Group on Oilseeds, Oils and Fats (30th Session), Grains (32nd Session) and Rice (43rd Session)*, University of Concepcion, Chile, and University of California, Berkeley.

Evans, H.J. & Wood, H.G. (1968). The mechanism of the pyruvate, phosphate dikinase reaction. *Proceedings of the National Academy of Sciences of the United States of America,* **61,** 1448-1453.

Fadrosh, D., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. & Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome,* **2,** 6.

Feria, A.-B., Alvarez, R., Cochereau, L., Vidal, J., García-Mauriño, S. & Echevarría, C. (2008). Regulation of Phosphoenolpyruvate Carboxylase Phosphorylation by Metabolites and

Abscisic Acid during the Development and Germination of Barley Seeds. *Plant Physiology,* **148,** 761-774.

Filichkin, S., Priest, H.D., Megraw, M. & Mockler, T.C. (2015). Alternative splicing in plants: directing traffic at the crossroads of adaptation and environmental stress. *Current opinion in plant biology,* **24,** 125-135.

Fontaine, V., Hartwell, J., Jenkins, G.I. & Nimmo, H.G. (2002). Arabidopsis thaliana contains two phosphoenolpyruvate carboxylase kinase genes with different expression patterns. *Plant, Cell & Environment,* **25,** 115-122.

Fowler, S., Lee, K., Onouchi, H., Samach, A., Richardson, K., Morris, B., Coupland, G. & Putterill, J. (1999). GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in Arabidopsis and encodes a protein with several possible membrane-spanning domains. *The EMBO Journal,* **18,** 4679-4688.

Franco, A., Ball, E. & Lüttge, U. (1990). Patterns of gas exchange and organic acid oscillations in tropical trees of the genus Clusia. *Oecologia,* **85,** 108-114.

Freeling, M. (1992). A conceptual framework for maize leaf development. *Dev Biol,* **153,** 44-58.

French, A.D. (1989). Chemical and physical properties of fructans. *Journal of plant physiology,* **134,** 125-136.

Freschi, L., Rodrigues, M.A., Tiné, M.a.S. & Mercier, H. (2010). Correlation between citric acid and nitrate metabolisms during CAM cycle in the atmospheric bromeliad Tillandsia pohliana. *Journal of plant physiology,* **167,** 1577-1583.

Fukayama, H., Tamai, T., Taniguchi, Y., Sullivan, S., Miyao, M. & Nimmo, H.G. (2006). Characterization and functional analysis of phosphoenolpyruvate carboxylase kinase genes in rice. *The Plant Journal,* **47,** 258-268.

Gachon, C., Mingam, A. & Charrier, B. (2004). Real-time PCR: what relevance to plant studies? *Journal of Experimental Botany,* **55,** 1445-1454.

García Mendoza, A. (2007). Los agaves de México. *Ciencias*.

Gentry, H.S. (2004). *Agaves of Continental North America,* University of Arizona Press.

Gerland, P., Raftery, A.E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B.K., Chunn, J. & Lalic, N. (2014). World population stabilization unlikely this century. *Science,* **346,** 234-237.

González, M.a.-C., EchevarríA, C., Vidal, J. & Cejudo, F.J. (2002). Isolation and characterisation of a wheat phosphoenolpyruvate carboxylase gene. Modelling of the encoded protein. *Plant Science,* **162,** 233-238.

Good-Avila, S.V., Souza, V., Gaut, B.S. & Eguiarte, L.E. (2006). Timing and rate of speciation in Agave (Agavaceae). *Proceedings of the National Academy of Sciences,* **103,** 9124-9129.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology,* **29,** 644-652.

Graf, A., Schlereth, A., Stitt, M. & Smith, A.M. (2010). Circadian control of carbohydrate availability for growth in Arabidopsis plants at night. *Proceedings of the National Academy of Sciences,* **107,** 9458-9463.

Greenham, K. & Mcclung, C.R. (2015). Integrating circadian dynamics with physiological processes in plants. *Nature Reviews Genetics,* **16,** 598-610.

Griffiths, H., Broadmeadow, M.J., Borland, A. & Hetherington, C. (1990). Short-term changes in carbon-isotope discrimination identify transitions between C3 and C4 carboxylation during Crassulacean acid metabolism. *Planta,* **181,** 604-610.

Gross, S.M., Martin, J.A., Simpson, J., Abraham-Juarez, M.J., Wang, Z. & Visel, A. (2013). De novo transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana. *BMC Genomics,* **14,** 563.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* **29,** 1072-1075.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N. & Regev, A. (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature protocols,* **8,** 10.1038/nprot.2013.1084.

Haas, B.J. & Zody, M.C. (2010). Advancing RNA-Seq analysis. *Nature biotechnology,* **28,** 421-423.

Hafke, J.B., Hafke, Y., Smith, J.A., Luttge, U. & Thiel, G. (2003). Vacuolar malate uptake is mediated by an anion-selective inward rectifier. *The Plant Journal,* **35,** 116-128.

Hake, S., Smith, H.M., Holtan, H., Magnani, E., Mele, G. & Ramirez, J. (2004). The role of knox genes in plant development. *Annu Rev Cell Dev Biol,* **20,** 125-151.

Harmer, S.L. (2009). The circadian system in higher plants. *Annual Review of Plant Biology,* **60,** 357-377.

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.-S., Han, B., Zhu, T., Wang, X., Kreps, J.A. & Kay, S.A. (2000). Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. *Science,* **290,** 2110-2113.

Hartsock, T.L. & Nobel, P.S. (1976). Watering converts a CAM plant to daytime CO2 uptake. *Nature,* **262,** 574-576.

Hartwell, J. (2005). The co-ordination of central plant metabolism by the circadian clock. *Biochem Soc Trans,* **33,** 945-948.

Hartwell, J. (2006). The Circadian Clock in CAM Plants. In: *Annual Plant Reviews Volume 21: Endogenous Plant Rhythms* (ed. by A. Hall & H. Mcwatters), pp. 211-236. Blackwell Publishing Ltd.

Hartwell, J., Gill, A., Nimmo, G.A., Wilkins, M.B., Jenkins, G.I. & Nimmo, H.G. (1999). Phosphoenolpyruvate carboxylase kinase is a novel protein kinase regulated at the level of expression. *The Plant Journal,* **20,** 333-342.

Hartwell, J., Smith, L.H., Wilkins, M.B., Jenkins, G.I. & Nimmo, H.G. (1996). Higher plant phosphoenolpyruvate carboxylase kinase is regulated at the level of translatable mRNA in response to light or a circadian rhythm. *The Plant Journal,* **10,** 1071-1078.

Häusler, R.E., Baur, B., Scharte, J., Teichmann, T., Eicks, M., Fischer, K.L., Flügge, U.-I., Schubert, S., Weber, A. & Fischer, K. (2000). Plastidic metabolite transporters and their physiological functions in the inducible crassulacean acid metabolism plant Mesembryanthemum crystallinum. *The Plant Journal,* **24,** 285-296.

Hay, A. & Tsiantis, M. (2009). A KNOX family TALE. *Current opinion in plant biology,* **12,** 593-598.

Haydon, M.J., Bell, L.J. & Webb, A.a.R. (2011). Interactions between plant circadian clocks and solute transport. *Journal of Experimental Botany,* **62,** 2333 - 2348.

Haydon, M.J., Mielczarek, O., Robertson, F.C., Hubbard, K.E. & Webb, A.A. (2013). Photosynthetic entrainment of the Arabidopsis thaliana circadian clock. *Nature,* **502,** 689-692.

Hendry, G.A. (1993). Evolutionary origins and natural functions of fructans-a climatological, biogeographic and mechanistic appraisal. *New Phytologist***,** 3-14.

Hennessey, T.L. & Field, C.B. (1991). Circadian rhythms in photosynthesis oscillations in carbon assimilation and stomatal conductance under constant conditions. *Plant Physiology,* **96,** 831-836.

Hodkinson, B.P. & Grice, E.A. (2015). Next-generation sequencing: a review of technologies and tools for wound microbiome research. *Advances in wound care,* **4,** 50-58.

Holtum, J., Winter, K. & Osmond, B. (1999). Crassulacean acid metabolism (CAM). In: *Plants in Action: Adaptation in Nature, Performance in Cultivation* (ed. by B.J. Atwell, P.E. Kriedemann & C.G.N. Turnbull). Macmillan Education Australia.

Holtum, J.A. & Winter, K. (2014). Limited photosynthetic plasticity in the leaf-succulent CAM plant Agave angustifolia grown at different temperatures. *Functional Plant Biology,* **41,** 843-849.

Holtum, J.A., Winter, K., Weeks, M.A. & Sexton, T.R. (2007). Crassulacean acid metabolism in the ZZ plant, Zamioculcas zamiifolia (Araceae). *Am J Bot,* **94,** 1670-1676.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. & Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Research,* **35,** W585-W587.

Hurst, A.C., Grams, T.E.E. & Ratajczak, R. (2004). Effects of salinity, high irradiance, ozone, and ethylene on mode of photosynthesis, oxidative stress and oxidative damage in the C3/CAM intermediate plant Mesembryanthemum crystallinum L. *Plant, Cell & Environment,* **27,** 187-197.

Hurth, M.A., Suh, S.J., Kretzschmar, T., Geis, T., Bregante, M., Gambale, F., Martinoia, E. & Neuhaus, H.E. (2005). Impaired pH Homeostasis in Arabidopsis Lacking the Vacuolar Dicarboxylate Transporter and Analysis of Carboxylic Acid Transport across the Tonoplast. *Plant Physiology,* **137,** 901-910.

Imaizumi, T. & Kay, S.A. (2006). Photoperiodic control of flowering: not only by coincidence. *Trends in Plant Science,* **11,** 550-558.

James, A.B., Monreal, J.A., Nimmo, G.A., Kelly, C.L., Herzyk, P., Jenkins, G.I. & Nimmo, H.G. (2008). The circadian clock in Arabidopsis roots is a simplified slave version of the clock in shoots. *Science,* **322,** 1832-1835.

James, A.B., Syed, N.H., Bordage, S., Marshall, J., Nimmo, G.A., Jenkins, G.I., Herzyk, P., Brown, J.W. & Nimmo, H.G. (2012). Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *The Plant Cell,* **24,** 961-981.

Jones, M.B. (1975). The effect of leaf age on leaf resistance and CO2 exchange of the CAM plant Bryophyllum fedtschenkoi. *Planta,* **123,** 91-96.

Joshi Na, F.J. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]* [Online]. Available: https://github.com/najoshi/sickle 2012].

Jouve, L., Greppin, H. & Agosti, R.D. (1998). Arabidopsis thaliana floral stem elongation: Evidence for an endogenous circadian rhythm. *Plant Physiology and Biochemistry,* **36,** 469-472.

Jung, B., Ludewig, F., Schulz, A., Meißner, G., Wöstefeld, N., Flügge, U.-I., Pommerrenig, B., Wirsching, P., Sauer, N., Koch, W., Sommer, F., Mühlhaus, T., Schroda, M., Cuin, T.A., Graus, D., Marten, I., Hedrich, R. & Neuhaus, H.E. (2015). Identification of the transporter responsible for sucrose accumulation in sugar beet taproots. *Nature Plants,* **1,** 14001.

Kahn, R.A. & Durst, F. (2000). Function and evolution of plant cytochrome P450. *Recent advances in phytochemistry,* **34,** 151-190.

Karin, M. (1990). Too many transcription factors: positive and negative interactions. *New Biology,* **2,** 126-131.

Kawakami, A. & Yoshida, M. (2002). Molecular characterization of sucrose:sucrose 1-fructosyltransferase and sucrose:fructan 6-fructosyltransferase associated with fructan

accumulation in winter wheat during cold hardening. *Biosci Biotechnol Biochem,* **66,** 2297-2305.

Keeley, J. (2014). Aquatic CAM photosynthesis: a brief history of its discovery. *Aquatic Botany,* **118,** 38-44.

Kendall, H.W. & Pimentel, D. (1994). Constraints on the expansion of the global food supply. *Ambio***,** 198-205.

Kenyon, W.H., Severson, R.F. & Black, C.C. (1985). Maintenance Carbon Cycle in Crassulacean Acid Metabolism Plant Leaves : Source and Compartmentation of Carbon for Nocturnal Malate Synthesis. *Plant Physiology,* **77,** 183-189.

Khan, A. (2001). *Plant anatomy and physiology,* Gyan Publishing House.

Kluge, M., Razanoelisoa, B. & Brulfert, J. (2001). Implications of Genotypic Diversity and Phenotypic Plasticity in the Ecophysiological Success of CAM Plants, Examined by Studies on the Vegetation of Madagascar1. *Plant Biology,* **3,** 214-222.

Knight, H., Thomson, A.J.W. & Mcwatters, H.G. (2008). Sensitive to freezing6 Integrates Cellular and Environmental Inputs to the Plant Circadian Clock. *Plant Physiology,* **148,** 293-303.

Kondo, A., Nose, A. & Ueno, O. (1998). Leaf inner structure and immunogold localization of some key enzymes involved in carbon metabolism in CAM plants. *Journal of Experimental Botany,* **49,** 1953-1961.

Kovermann, P., Meyer, S., Hörtensteiner, S., Picco, C., Scholz-Starke, J., Ravera, S., Lee, Y. & Martinoia, E. (2007). The Arabidopsis vacuolar malate channel is a member of the ALMT family. *The Plant Journal,* **52,** 1169-1180.

Krivorotova, T. & Sereikaite, J. (2014). Determination of fructan exohydrolase activity in the crude extracts of plants. *Electronic Journal of Biotechnology,* **17,** 329-333.

Laemmli, U.K. (1970). Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature,* **227,** 680-685.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A. & Huala, E. (2011). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*.

Lange, V., Bohme, I., Hofmann, J., Lang, K., Sauter, J., Schone, B., Paul, P., Albrecht, V., Andreas, J., Baier, D., Nething, J., Ehninger, U., Schwarzelt, C., Pingel, J., Ehninger, G. & Schmidt, A. (2014). Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics,* **15,** 63.

Langmead, B. & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods,* **9,** 357-359.

Latchman, D.S. (1997). Transcription factors: an overview. *The International Journal of Biochemistry & Cell Biology,* **29,** 1305-1312.

Le Roy, K., Lammens, W., Verhaest, M., De Coninck, B., Rabijns, A., Van Laere, A. & Van Den Ende, W. (2007). Unraveling the Difference between Invertases and Fructan Exohydrolases: A Single Amino Acid (Asp-239) Substitution Transforms Arabidopsis Cell Wall Invertase1 into a Fructan 1-Exohydrolase. *Plant Physiology,* **145,** 616-625.

Lee, D., Ainbinder, L., Erdembileg, S., Hurley, A., Morrison, L., Roig, M., Sibanda, A. & Yang, W. (2011). The global food crises. In: *The Global Social Crisis: Report on the World Social Situation 2011*. Department of Economic and Social Affairs, United Nations, New York.

Lehti-Shiu, M.D. & Shiu, S.H. (2012). Diversity, classification and function of the plant protein kinase superfamily. *Philos Trans R Soc Lond B Biol Sci,* **367,** 2619-2639.

Li, B. & Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics,* **12,** 323.

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R. & Dewey, C.N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol,* **15,** 553.

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S.L., Kebrom, T.H., Provart, N., Patel, R., Myers, C.R., Reidel, E.J., Turgeon, R., Liu, P., Sun, Q., Nelson, T. & Brutnell, T.P. (2010). The developmental dynamics of the maize leaf transcriptome. *Nature genetics,* **42,** 1060-1067.

Lister, R., O'malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. & Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell,* **133,** 523-536.

Livak, K.J. & Schmittgen, T.D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods,* **25,** 402-408.

Lopez, M.G., Mancilla-Margalli, N.A. & Mendoza-Diaz, G. (2003). Molecular structures of fructans from Agave tequilana Weber var. azul. *J Agric Food Chem,* **51,** 7835-7840.

Lujan, R., Lledias, F., Martinez, L.M., Barreto, R., Cassab, G.I. & Nieto-Sotelo, J. (2009). Small heat-shock proteins and leaf cooling capacity account for the unusual heat tolerance of the central spike leaves in Agave tequilana var. Weber. *Plant, Cell & Environment,* **32,** 1791-1803.

Lüttge, U. (1996). Clusia: Plasticity and Diversity in a Genus of C3/CAM Intermediate Tropical Trees. In: *Crassulacean Acid Metabolism* (ed. by K. Winter & J.A. Smith), pp. 296-311. Springer Berlin Heidelberg.

Lüttge, U. (2000). The tonoplast functioning as the master switch for circadian regulation of crassulacean acid metabolism. *Planta,* **211,** 761-769.

Lüttge, U. (2002). CO2-concentrating: consequences in crassulacean acid metabolism. *Journal of Experimental Botany,* **53,** 2131-2142.

Lüttge, U. (2004). Ecophysiology of crassulacean acid metabolism (CAM). *Annals of Botany,* **93,** 629-652.

Magrane, M. & Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database,* **2011,** bar009.

Maldonado-Sanchez, A.E. (2009). *Improved Agave Cultivars (Agave angustifolia Haw) for Profitable and Sustainable Bioethanol Production in Mexico* [Online]. Texcoco, Mexico: Chapingo Autonomous University. Available: http://www.nodai.ac.jp/cip/iss/english/9th_iss/fullpaper/2-1-3uach-maldonado.pdf [Accessed 11 November 2011].

Mallona, I., Egea-Cortines, M. & Weiss, J. (2011). Conserved and divergent rhythms of crassulacean acid metabolism-related and core clock gene expression in the cactus Opuntia ficus-indica. *Plant Physiology,* **156,** 1978-1989.

Mancilla-Margalli, N.A. & Lopez, M.G. (2006). Water-soluble carbohydrates and fructan structure patterns from Agave and Dasylirion species. *J Agric Food Chem,* **54,** 7832-7839.

Martin, J.A. & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics,* **12,** 671-682.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011,* **17**.

Martinoia, E., Maeshima, M. & Neuhaus, H.E. (2007). Vacuolar transporters and their essential role in plant metabolism. *Journal of Experimental Botany,* **58,** 83-102.

Matiz, A., Tamaso, P., Yepes, A., Freschi, L. & Mercier, H. (2013). CAM Photosynthesis in Bromeliads and Agaves: What Can We Learn from These Plants? In: *Photosynthesis*, pp. 91-134. InTech, Rijeka, Croatia.

Mcclung, C.R. (2006). Plant Circadian Rhythms. *The Plant Cell,* **18,** 792-803.

Mckain, M.R., Wickett, N., Zhang, Y., Ayyampalayam, S., Mccombie, W.R., Chase, M.W., Pires, J.C. & Leebens-Mack, J. (2012). Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *Am J Bot,* **99,** 397-406.

Mcrae, S.R., Christopher, J.T., Smith, J.a.C. & Holtum, J.A. (2002). Sucrose transport across the vacuolar membrane of Ananas comosus. *Functional Plant Biology,* **29,** 717-724.

Mellado-Mojica, E. & Lopez, M.G. (2012). Fructan metabolism in A. tequilana Weber Blue variety along its developmental cycle in the field. *J Agric Food Chem,* **60,** 11704-11713.

Mendicino, J. (1960). Sucrose phosphate synthesis in wheat germ and green leaves. *Journal of Biological Chemistry,* **235,** 3347-3352.

Michael, T.P. & Jackson, S. (2013). The First 50 Plant Genomes. *The Plant Genome,* **6**.

Michael, T.P. & Vanburen, R. (2015). Progress, challenges and the future of crop genomes. *Current opinion in plant biology,* **24,** 71-81.

Mielenz, J.R., Rodriguez, M., Jr., Thompson, O.A., Yang, X. & Yin, H. (2015). Development of Agave as a dedicated biomass source: production of biofuels from whole plants. *Biotechnol Biofuels,* **8,** 79.

Miller, J.R., Koren, S. & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics,* **95,** 315-327.

Möllering, H. (1974). L-Malate: determination with malate dehydrogenase and glutamate-oxaloacetate transaminase. In: *BergmeyerHU, ed. Methods of enzymatic analysis*, pp. 1589-1593. Verlag Chemie, Weinheim, Germany

Morant, M., Bak, S., Møller, B.L. & Werck-Reichhart, D. (2003). Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. *Current Opinion in Biotechnology,* **14,** 151-162.

Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T.J., Mcdonald, H., Varhol, R., Jones, S.J. & Marra, M.A. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques,* **45,** 81.

Morozova, O. & Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics,* **92,** 255-264.

Mortazavi, A., Williams, B.A., Mccue, K., Schaeffer, L. & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods,* **5,** 621-628.

Morvan-Bertrand, A., Boucaud, J. & Prud'homme, M.-P. (1999). Influence of initial levels of carbohydrates, fructans, nitrogen, and soluble proteins on regrowth of Lolium perenne L. cv. Bravo following defoliation. *Journal of Experimental Botany,* **50,** 1817-1826.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science,* **320,** 1344-1349.

Nimmo, G., Wilkins, M., Fewson, C. & Nimmo, H. (1987). Persistent circadian rhythms in the phosphorylation state of phosphoenolpyruvate carboxylase from Bryophyllum fedtschenkoi leaves and in its sensitivity to inhibition by malate. *Planta,* **170,** 408-415.

Nimmo, G.A., Nimmo, H., Fewson, C. & Wilkins, M. (1984). Diurnal changes in the properties of phosphoenolpyruvate carboxylase in Bryophyllum leaves: a possible co valent modification. *FEBS letters,* **178,** 199-203.

Nimmo, G.A., Nimmo, H., Hamilton, I.D., Fewson, C.A. & Wilkins, M.B. (1986). Purification of the phosphorylated night form and dephosphorylated day form of phosphoenolpyruvate carboxylase from Bryophyllum fedtschenkoi. *Biochemical Journal,* **239,** 213-220.

Nimmo, H.G. (2000). The regulation of phosphoenolpyruvate carboxylase in CAM plants. *Trends in Plant Science,* **5,** 75-80.

Nobel, P. (2010). *Desert wisdom/agaves and cacti: CO2, water, climate change. iUniverse,* Inc., New York.

Nobel, P.S. (1985). Water Relations and Carbon Dioxide Uptake of Agave deserti-Special Adaptations to Desert Climates. *Desert plants*.

Nobel, P.S. (1994). *Remarkable agaves and cacti,* Oxford University Press.

Nobel, P.S. (1996). High Productivity of Certain Agronomic CAM Species. In: *Crassulacean Acid Metabolism* (ed. by K. Winter & J.A. Smith), pp. 255-265. Springer Berlin Heidelberg.

Nobel, P.S. & Sanderson, J. (1984). Rectifier-like activities of roots of two desert succulents. *Journal of Experimental Botany,* **35,** 727-737.

Nobel, P.S. & Valenzuela, A.G. (1987). Environmental responses and productivity of the CAM plant, Agave tequilana. *Agricultural and Forest Meteorology,* **39,** 319-334.

North, G. & Nobel, P. (1998). Water uptake and structural plasticity along roots of a desert succulent during prolonged drought. *Plant, Cell & Environment,* **21,** 705-713.

Ogden, R.C. & Adams, D.A. (1987). Electrophoresis in agarose and acrylamide gels. *Methods Enzymol,* **152,** 61-87.

Olivares, E. & Medina, E. (1990). Carbon dioxide exchange, soluble carbohydrates and acid accumulation in a fructan accumulating plant: Fourcroya humboldtiana Treal. *Journal of Experimental Botany,* **41,** 579-585.

Oliveros, J.C. (2007-2015). *Venny. An interactive tool for comparing lists with Venn's diagrams* [Online]. Available: http://bioinfogp.cnb.csic.es/tools/venny/index.html [Accessed 10 October 2014].

Osmond, C.B. (1978). Crassulacean Acid Metabolism: A Curiosity in Context. *Annual Review of Plant Physiology,* **29,** 379-414.

Palmieri, L., Picault, N., Arrigoni, R., Besin, E., Palmieri, F. & Hodges, M. (2008). Molecular identification of three Arabidopsis thaliana mitochondrial dicarboxylate carrier isoforms: organ distribution, bacterial expression, reconstitution into liposomes and functional characterization. *Biochemical Journal,* **410,** 621-629.

Palomino, G., Dolezel, J., Méndez, I. & Rubluo, A. (2003). Nuclear genome size analysis of Agave tequilana Weber. *Caryologia,* **56,** 37-46.

Park, D.H. (1999). Control of Circadian Rhythms and Photoperiodic Flowering by the Arabidopsis GIGANTEA Gene. *Science,* **285,** 1579-1582.

Parra, G., Bradnam, K. & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics,* **23,** 1061-1067.

Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. (2009). Assessing the gene space in draft genomes. *Nucleic Acids Research,* **37,** 289-297.

Pfaffl, M.W., Tichopad, A., Prgomet, C. & Neuvians, T.P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. *Biotechnol Lett,* **26,** 509-515.

Picault, N., Palmieri, L., Pisano, I., Hodges, M. & Palmieri, F. (2002). Identification of a novel transporter for dicarboxylates and tricarboxylates in plant mitochondria. Bacterial expression, reconstitution, functional characterization, and tissue distribution. *Journal of Biological Chemistry,* **277,** 24204-24211.

Pilon-Smits, E.a.H., Hart, H.T. & Van Brederode, J. (1996). Evolutionary Aspects of Crassulacean Acid Metabolism in the Crassulaceae. In: *Crassulacean Acid Metabolism* (ed. by K. Winter & J.A. Smith), pp. 349-359. Springer Berlin Heidelberg.

Potenza, E., Racchi, M., Sterck, L., Coller, E., Asquini, E., Tosatto, S., Velasco, R., Van De Peer, Y. & Cestaro, A. (2015). Exploration of alternative splicing events in ten different grapevine cultivars. *BMC Genomics,* **16,** 706.

Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. & Finn, R. (2012). The Pfam protein families database *Nucleic Acids Research,* **40,** D290-D301.

Quilliam, R.S., Swarbrick, P.J., Scholes, J.D. & Rolfe, S.A. (2006). Imaging photosynthesis in wounded leaves of Arabidopsis thaliana. *Journal of Experimental Botany,* **57,** 55-69.

Ragauskas, A.J., Williams, C.K., Davison, B.H., Britovsek, G., Cairney, J., Eckert, C.A., Frederick, W.J., Hallett, J.P., Leak, D.J., Liotta, C.L., Mielenz, J.R., Murphy, R., Templer, R. & Tschaplinski, T. (2006). The Path Forward for Biofuels and Biomaterials. *Science,* **311,** 484-489.

Raveh, E., Wang, N. & Nobel, P.S. (1998). Gas exchange and metabolite fluctuations in green and yellow bands of variegated leaves of the monocotyledonous CAM species Agave americana. *Physiologia Plantarum,* **103,** 99-106.

Raven, J.A. & Spicer, R.A. (1996). The Evolution of Crassulacean Acid Metabolism. In: *Crassulacean Acid Metabolism* (ed. by K. Winter & J.A. Smith), pp. 360-385. Springer Berlin Heidelberg.

Ray, W.J. & Roscelli, G.A. (1964). A Kinetic Study of the Phosphoglucomutase Pathway. *Journal of Biological Chemistry,* **239,** 1228-1236.

Reddy, A.S.N., Marquez, Y., Kalyna, M. & Barta, A. (2013). Complexity of the Alternative Splicing Landscape in Plants. *The Plant Cell,* **25,** 3657-3683.

Ritsema, T. & Smeekens, S. (2003). Fructans: beneficial for plants and humans. *Current opinion in plant biology,* **6,** 223-230.

Robinson, M.D., Mccarthy, D.J. & Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* **26,** 139-140.

Rolland, F., Moore, B. & Sheen, J. (2002). Sugar Sensing and Signaling in Plants. *The Plant Cell,* **14,** S185-S205.

Salomé, P.A. & Mcclung, C.R. (2005). PSEUDO-RESPONSE REGULATOR 7 and 9 are partially redundant genes essential for the temperature responsiveness of the Arabidopsis circadian clock. *The Plant Cell,* **17,** 791-803.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., Marcais, G., Pop, M. & Yorke, J.A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res,* **22,** 557-567.

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* **270,** 467-470.

Scofield, S., Dewitte, W., Nieuwland, J. & Murray, J.A. (2013). The Arabidopsis homeobox gene SHOOT MERISTEMLESS has cellular and meristem-organisational roles with differential requirements for cytokinin and CYCD3 activity. *The Plant Journal,* **75,** 53-66.

Shenton, M., Fontaine, V., Hartwell, J., Marsh, J.T., Jenkins, G.I. & Nimmo, H.G. (2006). Distinct patterns of control and expression amongst members of the PEP carboxylase kinase gene family in C4 plants. *The Plant Journal,* **48,** 45-53.

Shreve, F. (1942). The desert vegetation of North America. *The Botanical Review,* **8,** 195-246.

Silvera, K., Neubig, K.M., Whitten, W.M., Williams, N.H., Winter, K. & Cushman, J.C. (2010). Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology,* **37,** 995-1010.

Singh, S., Sundaram, S. & Kishor, K. (2014). Carbon-Concentrating Mechanism. In: *Photosynthetic Microorganisms*, pp. 5-38. Springer International Publishing.

Small, I., Peeters, N., Legeai, F. & Lurin, C. (2004). Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *PROTEOMICS,* **4,** 1581-1590.

Smith, J.a.C. & Bryce, J.H. (1992). Metabolite compartmentation and transport in CAM plants Plant In: *Organelles Compartmentation of Metabolism in Photosynthetic Cells*. Cambridge University Press.

Smith, S., Monson, R. & Anderson, J. (1997). CAM Succulents. In: *Physiological Ecology of North American Desert Plants*, pp. 125-140. Springer Berlin Heidelberg.

Somers, D.E., Webb, A., Pearson, M. & Kay, S.A. (1998). The short-period mutant, toc1-1, alters circadian clock regulation of multiple outputs throughout development in Arabidopsis thaliana. *Development,* **125,** 485-494.

Somerville, C., Youngs, H., Taylor, C., Davis, S.C. & Long, S.P. (2010). Feedstocks for Lignocellulosic Biofuels. *Science,* **329,** 790-792.

Steudle, E., Smith, J.a.C. & Lüttge, U. (1980). Water-relation parameters of individual mesophyll cells of the crassulacean acid metabolism plant Kalanchoë daigremontiana. *Plant Physiology,* **66,** 1155-1163.

Stewart, J.R. (2015). Agave as a model CAM crop system for a warming and drying world. *Frontiers in Plant Science,* **6,** 684.

Tauzin, A.S., Sulzenbacher, G., Lafond, M., Desseaux, V., Reca, I.B., Perrier, J., Bellincampi, D., Fourquet, P., Leveque, C. & Giardina, T. (2014). Functional characterization of a vacuolar invertase from Solanum lycopersicum: post-translational regulation by N-glycosylation and a proteinaceous inhibitor. *Biochimie,* **101,** 39-49.

Taybi, T., Nimmo, H.G. & Borland, A.M. (2004). Expression of phosphoenolpyruvate carboxylase and phosphoenolpyruvate carboxylase kinase genes. Implications for genotypic capacity and phenotypic plasticity in the expression of crassulacean acid metabolism. *Plant Physiology,* **135,** 587-598.

Taybi, T., Patil, S., Chollet, R. & Cushman, J.C. (2000). A minimal serine/threonine protein kinase circadianly regulates phosphoenolpyruvate carboxylase activity in crassulacean acid metabolism-induced leaves of the common ice plant. *Plant Physiology,* **123,** 1471-1482.

Theng, V., Agarie, S. & Nose, A. (2008). Regulatory phosphorylation of phosphoenolpyruvate carboxylase in the leaves of Kalanchoe pinnata, K. daigremontiana and Ananas comosus. *Biologia Plantarum,* **52,** 281-290.

Thomas, M. (1949). Physiological studies on acid metabolism in green plants. I. CO2 fixation and CO2 liberation in crassulacean acid metabolism. *New Phytologist,* **48,** 390-420.

Ting, I.P. (1985). Crassulacean Acid Metabolism. *Annual Review of Plant Physiology,* **36,** 595-622.

Tiwari, A., Kumar, P., Singh, S. & Ansari, S.A. (2005). Carbonic anhydrase in relation to higher plants. *Photosynthetica,* **43,** 1-11.

Topper, Y.J. (1957). On the mechanism of action of phosphoglucose isomerase and phosphomannose isomerase. *Journal of Biological Chemistry,* **225,** 419-426.

Tronconi, M.A., Fahnenstich, H., Gerrard Weehler, M.C., Andreo, C.S., Flügge, U.-I., Drincovich, M.F. & Maurino, V.G. (2008). Arabidopsis NAD-Malic Enzyme Functions As a Homodimer and Heterodimer and Has a Major Impact on Nocturnal Metabolism. *Plant Physiology,* **146,** 1540-1552.

Tsiantis, M.S., Bartholomew, D.M. & Smith, J.A. (1996). Salt regulation of transcript levels for the c subunit of a leaf vacuolar H(+)-ATPase in the halophyte Mesembryanthemum crystallinum. *The Plant Journal,* **9,** 729-736.

Tyagi, S. (2000). Taking a census of mRNA populations with microbeads. *Nature biotechnology,* **18,** 597-598.

Ueno, K., Ishiguro, Y., Yoshida, M., Onodera, S. & Shiomi, N. (2011). Cloning and functional characterization of a fructan 1-exohydrolase (1-FEH) in edible burdock (Arctium lappa L.). *Chemistry Central Journal,* **5,** 16-16.

Valluru, R. & Van Den Ende, W. (2008). Plant fructans in stress environments: emerging concepts and future prospects. *Journal of Experimental Botany,* **59,** 2905-2916.

Vargas-Ponce, O., Zizumbo-Villarreal, D. & Marin, P.C.-G. (2007). In situ diversity and maintanenace of traditional Agave landraces used in spirits production in West-Central Mexico. *Economic Botany,* **61,** 362-375.

Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. (1995). Serial analysis of gene expression. *Science,* **270,** 484-487.

Walters, R.G., Ibrahim, D.G., Horton, P. & Kruger, N.J. (2004). A Mutant of Arabidopsis Lacking the Triose-Phosphate/Phosphate Translocator Reveals Metabolic Regulation of Starch Breakdown in the Light. *Plant Physiology,* **135,** 891-906.

Wang, L., Li, X.R., Lian, H., Ni, D.A., He, Y.K., Chen, X.Y. & Ruan, Y.L. (2010). Evidence that high activity of vacuolar invertase is required for cotton fiber and Arabidopsis root elongation through osmotic dependent and independent pathways, respectively. *Plant Physiology,* **154,** 744-756.

Wang, N. & Nobel, P.S. (1998). Phloem Transport of Fructans in the Crassulacean Acid Metabolism Species Agave deserti. *Plant Physiology,* **116,** 709-714.

Wang, S.M. (2007). Understanding SAGE data. *Trends in Genetics,* **23,** 42-50.

Wang, Z.-Y. & Tobin, E.M. (1998). Constitutive Expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) Gene Disrupts Circadian Rhythms and Suppresses Its Own Expression. *Cell,* **93,** 1207-1217.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics,* **10,** 57-63.

Ward, J.M., Maser, P. & Schroeder, J.I. (2009). Plant ion channels: gene families, physiology, and functional genomics analyses. *Annu Rev Physiol,* **71,** 59-82.

Webb, A. (1998). Stomatal rhythms. In: *Biological rhythms and photoperiodism in plants* (ed. by P. Lumsden & A. Millar), pp. 69-79. Bios Scientific, Oxford, UK.

Weber, A.P., Weber, K.L., Carr, K., Wilkerson, C. & Ohlrogge, J.B. (2007). Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology,* **144,** 32-42.

Werck-Reichhart, D., Bak, S. & Paquette, S. (2002). Cytochromes P450. *The Arabidopsis book/American Society of Plant Biologists,* **1**.

Wilkins, M.B. (1959). An Endogenous Rhythm in the Rate of Carbon Dioxide Output of Bryophyllum: I. Some preliminary experiments. *Journal of Experimental Botany,* **10,** 377-390.

Wilkins, M.B. (1992). Circadian rhythms: their origin and control. *New Phytologist,* **121,** 347-375.

Winter, K. (1985). Crassulacean acid metabolism. In: *Photosynthetic Mechanisms and the Environment (Topics in photosynthesis), Vol. 6.* (ed. by J. Barber & N.R. Baker), pp. 329-387. Elsevier Biomedical Press, Amsterdam, Netherlands.

Winter, K., Aranda, J. & Holtum, J.a.M. (2005). Carbon isotope composition and water-use efficiency in plants with crassulacean acid metabolism. *Functional Plant Biology,* **32,** 381-388.

Winter, K. & Gademann, R. (1991). Daily Changes in CO(2) and Water Vapor Exchange, Chlorophyll Fluorescence, and Leaf Water Relations in the Halophyte Mesembryanthemum crystallinum during the Induction of Crassulacean Acid Metabolism in Response to High NaCl Salinity. *Plant Physiology,* **95,** 768-776.

Winter, K., Garcia, M. & Holtum, J.A. (2014). Nocturnal versus diurnal CO2 uptake: how flexible is Agave angustifolia? *Journal of Experimental Botany,* **65,** 3695-3703.

Winter, K., Holtum, J.A. & Smith, J.A. (2015). Crassulacean acid metabolism: a continuous or discrete trait? *New Phytologist,* **208,** 73–78.

Winter, K. & Holtum, J.a.M. (2011). Induction and reversal of crassulacean acid metabolism in Calandrinia polyandra: effects of soil moisture and nutrients. *Functional Plant Biology,* **38,** 576-582.

Winter, K. & Smith, J.a.C. (1996). An Introduction to Crassulacean Acid Metabolism. Biochemical Principles and Ecological Diversity. In: *Crassulacean Acid Metabolism* (ed. by K. Winter & J.A. Smith), pp. 1-13. Springer Berlin Heidelberg.

Wormit, A., Trentmann, O., Feifer, I., Lohr, C., Tjaden, J., Meyer, S., Schmidt, U., Martinoia, E. & Neuhaus, H.E. (2006). Molecular Identification and Physiological Characterization of a Novel Monosaccharide Transporter from Arabidopsis Involved in Vacuolar Sugar Transport. *The Plant Cell,* **18,** 3476-3490.

Yan, X., Tan, D.K., Inderwildi, O.R., Smith, J. & King, D.A. (2011). Life cycle energy and greenhouse gas analysis for agave-derived bioethanol. *Energy & Environmental Science,* **4,** 3110-3121.

Yang, X., Cushman, J.C., Borland, A.M., Edwards, E.J., Wullschleger, S.D., Tuskan, G.A., Owen, N.A., Griffiths, H., Smith, J.a.C., De Paoli, H.C., Weston, D.J., Cottingham, R., Hartwell, J., Davis, S.C., Silvera, K., Ming, R., Schlauch, K., Abraham, P., Stewart, J.R., Guo, H.-B., Albion, R., Ha, J., Lim, S.D., Wone, B.W.M., Yim, W.C., Garcia, T., Mayer, J.A., Petereit, J., Nair, S.S., Casey, E., Hettich, R.L., Ceusters, J., Ranjan, P., Palla, K.J., Yin, H., Reyes-García, C., Andrade, J.L., Freschi, L., Beltrán, J.D., Dever, L.V., Boxall, S.F., Waller, J., Davies, J., Bupphada, P., Kadu, N., Winter, K., Sage, R.F., Aguilar, C.N., Schmutz, J., Jenkins, J. & Holtum, J.a.M. (2015). A roadmap for research on crassulacean acid

metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist,* **207,** 491-504.

Yanovsky, M.J. & Kay, S.A. (2003). Living by the calendar: how plants know when to flower. *Nature Reviews Molecular Cell Biology,* **4,** 265-276.

Yu, C.P., Chen, S.C., Chang, Y.M., Liu, W.Y., Lin, H.H., Lin, J.J., Chen, H.J., Lu, Y.J., Wu, Y.H., Lu, M.Y., Lu, C.H., Shih, A.C., Ku, M.S., Shiu, S.H., Wu, S.H. & Li, W.H. (2015). Transcriptome dynamics of developing maize leaves and genomewide prediction of cis elements and their cognate transcription factors. *Proceedings of the National Academy of Sciences of the United States of America,* **112,** E2477-2486.

Zhou, W.Z., Zhang, Y.M., Lu, J.Y. & Li, J.F. (2012). Construction and evaluation of normalized cDNA libraries enriched with full-length sequences for rapid discovery of new genes from Sisal (Agave sisalana Perr.) different developmental stages. *International Journal of Molecular Sciences,* **13,** 13150-13168.

Appendix

Chapter 5 - Supplementary Data

5.1 Generation of the total RNA-samples, RNA QC and library production

Total RNA was isolated from the same *A. sisalana* leaf sections (white basal/ proximal tissue, pale green basal and dark green tip tissue sampled at 10:00 light and 22:00 dark) studied in chapters 3 and 4 including all three biological replicates. Each leaf section sample was ground to a fine powder in liquid nitrogen before sub-aliquots were used for the metabolite extracts, protein extracts and RNA isolations. Thus, the biochemical data in chapter 4 (malate, soluble sugars, and immuno-blot results for protein abundance) correlate directly with the RNA-seq data presented in this chapter. The 18 total RNA samples were submitted to the CGR, University of Liverpool and passed QC tests in terms of both quantity and quality. Bar-coded Illumina sequencing libraries were generated for each of the 18 total RNA samples by the staff in the CGR using the ScriptSeq Complete (plant leaf) kit (Epicentre) according to the manufacturer's protocols. The SCriptSeq Complete library kit includes a ribosomal RNA depletion kit specifically designed to capture plant ribosomal RNAs that would otherwise swamp the sequencing reads due to the high relative abundance of rRNA in total RNA samples

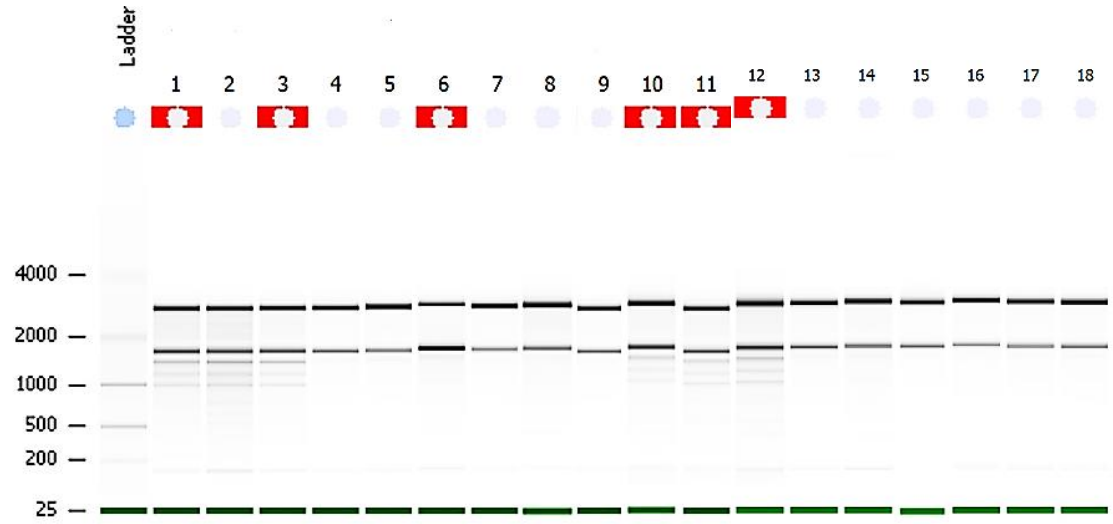
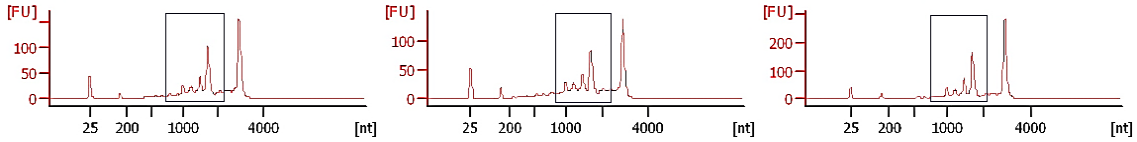
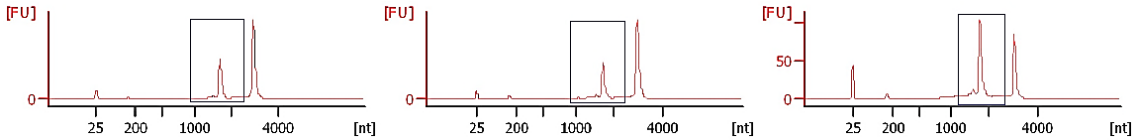


Figure S5.1 The corresponding gel-like image of RNA fragments of 18 samples generated using Aligent Bioanalyser 2100.

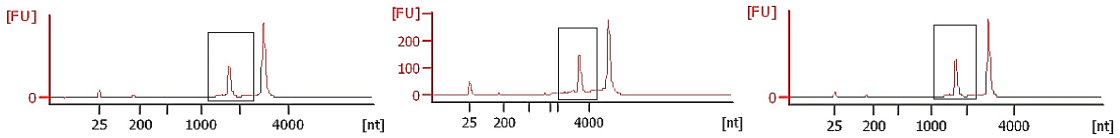
1-3 represent 3 biological replicates of the youngest fully expanded *A. sisalana* leaf tip, 4-6 : base and 7-9 : white part sampled at light (10 h), and 10-12 : tip, 13-15 : base and 16-18 : white part sampled at dark (22 h). The corresponding gel band-like image indicates RNAs quality intactness. The data was generated using RNA 6000 Pico Chip Kit run on Agilent Bioanalyzer equipment. The Y-axis represented the RNA size (nt).



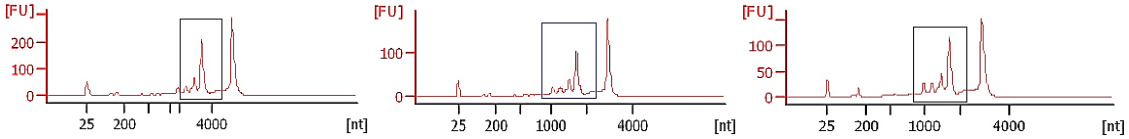
(A) Tip Light (10h) 3 biological replicates



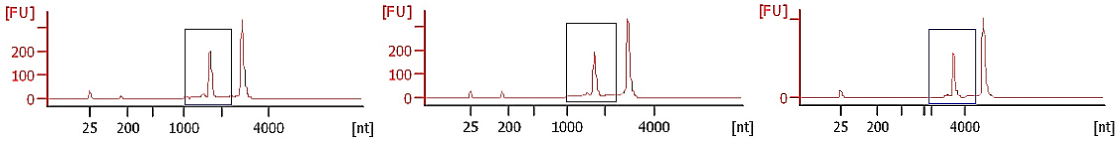
(B) Base Light (10h) 3 biological replicates



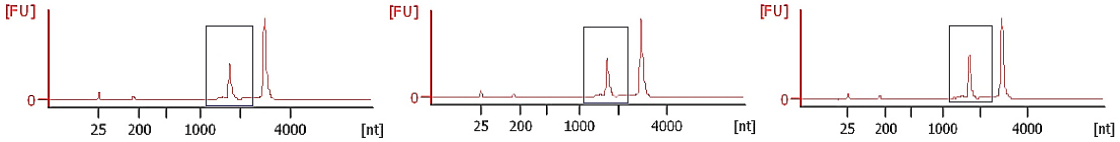
(C) White Light (10h) biological replicates



(D) Tip Dark (22h) 3 biological replicates



(E) Base Dark (22h) 3 biological replicates



(F) White Dark (22h) biological replicates

Figure S5.2 The RNA qualitative peaks of all 18 total samples produced from electropherograms using Agilent Bioanalyzer 2100 of all 18 total RNA samples.

1-3 represent 3 biological replicates of the youngest fully expanded *A. sisalana* leaf tip (A), 4-6 : base (B) and 7-9 : white part (C) sampled at light (10 h), and 10-12 : tip (D), 13-15 : base (E) and 16-18 : white part (F) sampled at dark (22 h). This indicates RNAs quality intactness. The data was generated using RNA 6000 Pico Chip Kit run on Agilent Bioanalyzer 2100 machine. The y-axis indicates fluorescence units and the x-axis indicates RNA length in nucleotides. The frame indicates the rRNA 18S peak and the peak on the right is 28S peak.

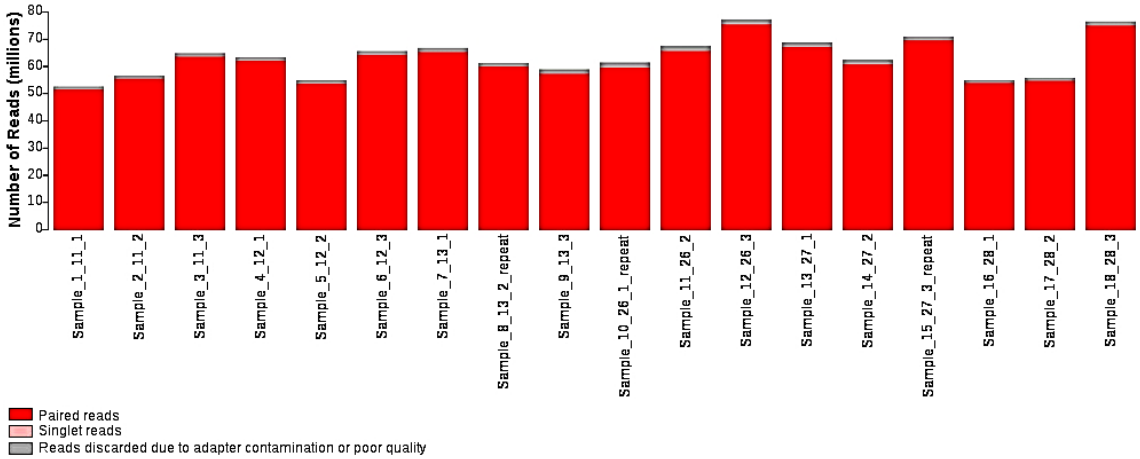


Figure S5.3 RNA-seq reads from all 18 samples

The numbers of reads per sample were recovered from Illumina HiSeq2500 sequencer. Sample 1-3 represent 3 biological replicates of the youngest fully expanded *A. sisalana* leaf tip, 4-6 : base and 7-9 : white part sampled at in the light (10:00), and 10-12 : tip, 13-15 : base and 16-18 : white part sampled in the dark (22:00).

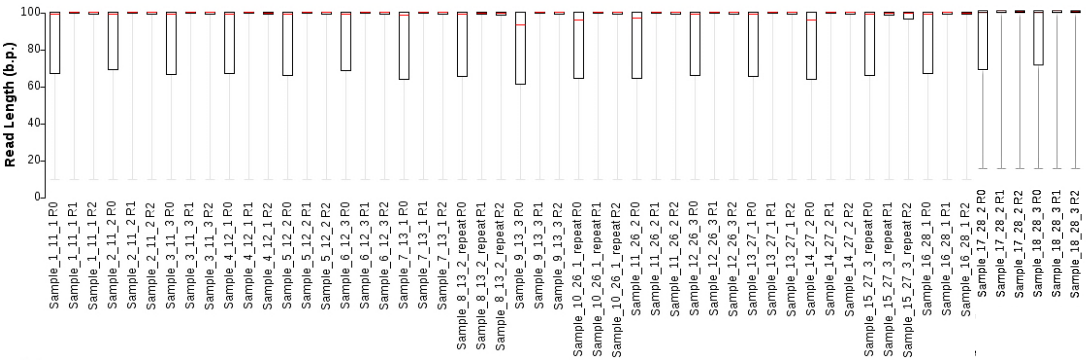


Figure S5.4 Read length of forward (R1), reverse (R2) and singlet (R0) of all 18 samples.

Red line indicates median length. Box indicates interquartile range. Whiskers indicate minimum and maximum read lengths. Box plot showing the distribution of trimmed read lengths for the forward (R1), reverse (R2) and singlet (R0) reads. Note that it is common for a small number of reads to consist of mostly adapter-derived sequence, so it is expected that the distribution will show a long tail. Sample 1-3 represent 3 biological replicates of the youngest fully expanded *A. sisalana* leaf tip, 4-6 : base and 7-9 : white part sampled at in the light (10:00 light), and 10-12 : tip, 13-15 : base and 16-18 : white part sampled in the dark (22:00 dark).

Table S5.1 Top 20 known CAM pathway genes that were found amongst the most highly expressed tip genes.

Contig name	Acronym	Gene name	Log-fold change Tip>Base	Log-fold change Tip>White	FDR Tip>Base	FDR Tip>White
c489202_g2	*As_PPC*	Phosphoenolpyruvate carboxylase	6.13	7.43	3.71E-84	2.04E-110
c582982_g3	*As_PPC*	Phosphoenolpyruvate carboxylase	5.82	7.24	5.47E-80	1.71E-106
c477309_g1	*As_PPCK*	Phosphoenolpyruvate carboxylase kinase	5.26	1.75	4.11E-15	0.001666
c525272_g3	*As_PPDK*	Pyruvate Orthophosphate dikinase	5.08	8.28	1.10E-52	4.38E-105
c502104_g1	*As_PPDK*	Pyruvate Orthophosphate dikinase	4.19	5.22	1.44E-47	2.48E-66
c520456_g1	*As_NAD-MDH*	NAD-malate dehydrogenase	3.67	4.53	5.33E-72	5.17E-99
c506574_g1	*As_ALMT*	Aluminium-activated malate transporter	3.36	4.79	1.96E-10	1.72E-16
c525409_g1	*As_NAD-ME*	NAD-malic enzyme	3.32	5.28	3.92E-15	5.80E-24
c565688_g2	*As_NAD-ME*	NAD-malic enzyme	2.80	3.39	5.27E-54	5.20E-76
c513606_g1	*As_NAD-ME*	NAD-malic enzyme	2.74	4.03	1.41E-34	7.69E-67
c500822_g1	*As_NAD-ME*	NAD-malic enzyme	2.61	3.56	1.41E-22	1.59E-38
c558023_g1	*As_ALMT*	Aluminium-activated malate transporter	2.49	3.50	6.04E-31	2.04E-56
c477860_g1	*As_ALMT*	Aluminium-activated malate transporter	2.24	5.16	6.76E-09	5.12E-22
c537822_g1	*As_NAD-ME*	NAD-malic enzyme	2.23	3.48	4.83E-21	1.19E-43
c583225_g1	*As_NAD-ME*	NAD-malic enzyme	2.04	2.35	1.19E-25	1.54E-33
c453533_g1	*As_NAD-ME*	NAD-malic enzyme	1.84	2.81	2.86E-20	3.44E-44
c561463_g1	*As_PPCK*	Phosphoenolpyruvate carboxylase kinase	1.82	3.30	2.13E-08	1.09E-24
c495715_g1	*As_NAD-ME*	NAD-malic enzyme	1.67	3.73	0.027113	2.40E-05
c564067_g1	*As_NAD-ME*	NAD-malic enzyme	1.57	2.64	6.91E-14	2.18E-36
c595219_g2	*As_ALMT*	Aluminium-activated malate transporter	1.50	2.11	5.60E-06	9.54E-10

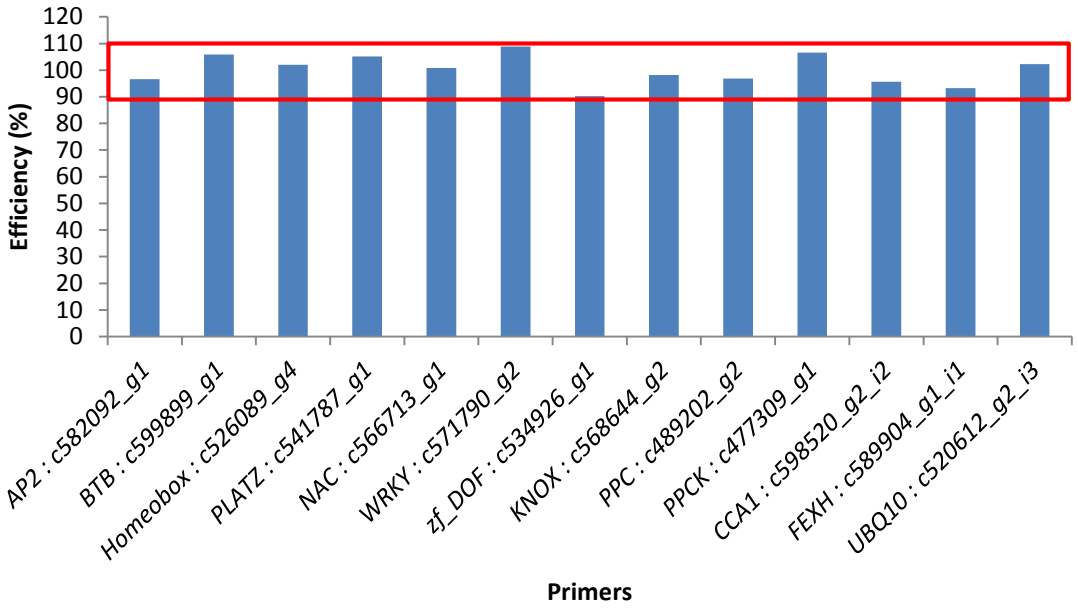
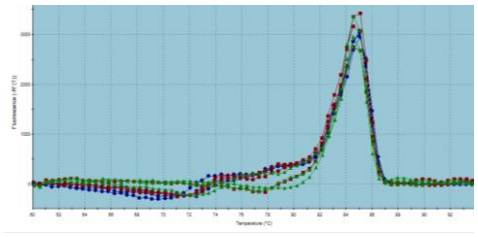


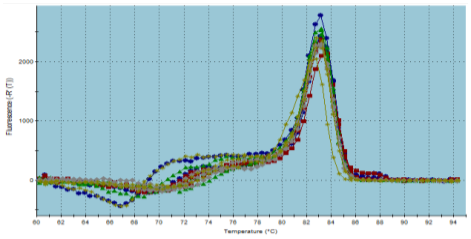
Figure S6.1 PCR reaction efficiency (%) of primers of newly discovered transcription factors, CAM control, circadian control, and reference genes.

PCR reaction efficiency percentage was calculated by MxPro 4.1 QPCR Software that came with the Mx3005P qPCR System. The y-axis indicates the primer efficiency. The x-axis indicates the primers. The red frame above the bar chart indicates the adequate primer efficiency (90-110%). All primers contain adequate primer efficiency.

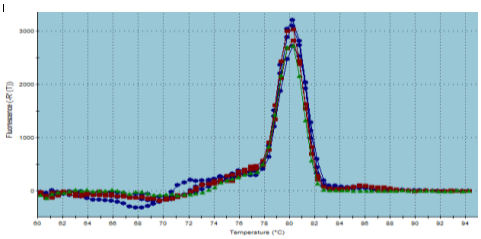
(A) *AP2* : c582092_g1



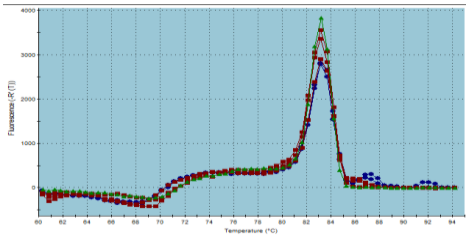
(B) *BTB* : c599899_g1



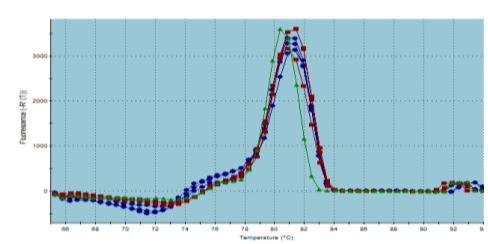
(C) *Homeobox* : c526089_g4



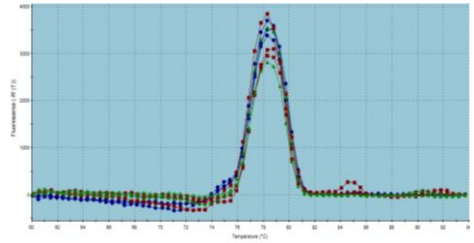
(D) *PLATZ* : c541787_g1



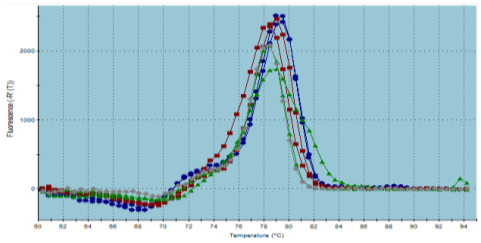
(E) *zf_DOF* : c534926_g1



(F) *NAC* : c566713_g1



(G) *WRKY* : c571790_g2



(H) *KNOX* : c568644_g2

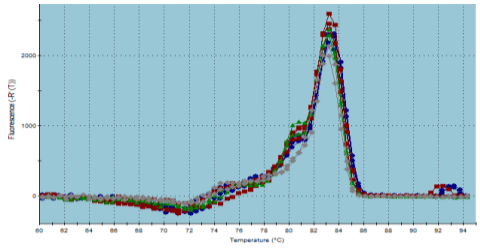
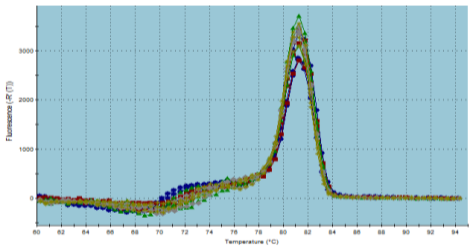


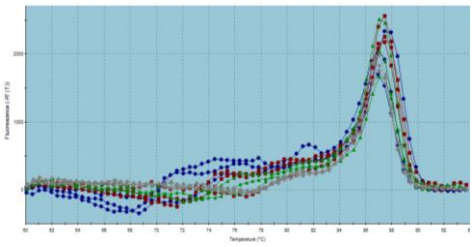
Figure S6.2 Dissociation (melting) curves of novel discovered CAM-induced and non-CAM genes.

Dissociation (melting) curves were analysed using MxPro 4.1 QPCR Software that came with the Mx3005P qPCR System. Novel discovered CAM-induced genes: *AP2* (A), *BTB* (B), *Homeobox* (C), *PLATZ* (D), *zf_DOF* (E), *NAC* (F), *WRKY* (G) and non-CAM gene: *KNOX* (H). The y-axis indicates changes in fluorescence level. The x-axis indicates temperature (C°).

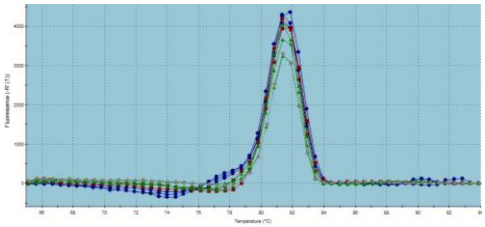
(A) *PPC* : c489202_g2



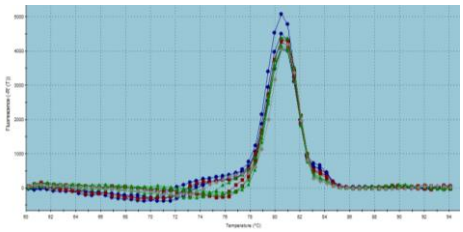
(B) *PPCK* : c477309_g1



(C) *CCA1* : c598520_g2_i2



(D) *FEXH* : c589904_g1_i1



(E) *UBQ10* : c520612_g2_i3

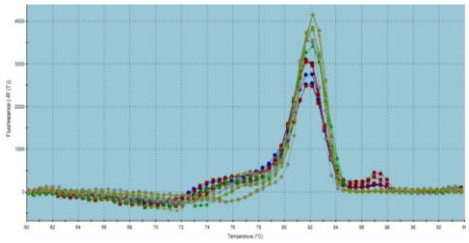


Figure S6.3 Dissociation (melting) curves of control CAM, circadian clock, and reference genes.

Dissociation (melting) curves were analysed using MxPro 4.1 QPCR Software that came with the Mx3005P qPCR System. Known CAM control genes: *PPC* (A) and *PPCK* (B), circadian clock control genes: *CCA1* (C) and *FEXH* (C), and reference gene: *UBQ10* (E). The y-axis indicates changes in fluorescence level. The x-axis indicates temperature (C°).